

Handbook^{of}

Teunissen

Montenbruck

Editors

 Springer

Springer Handbook of Global Navigation Satellite Systems

Springer Handbooks provide a concise compilation of approved key information on methods of research, general principles, and functional relationships in physical and applied sciences. The world's leading experts in the fields of physics and engineering will be assigned by one or several renowned editors to write the chapters comprising each volume. The content is selected by these experts from Springer sources (books, journals, online content) and other systematic and approved recent publications of scientific and technical information.

The volumes are designed to be useful as readable desk reference book to give a fast and comprehensive overview and easy retrieval of essential reliable key information, including tables, graphs, and bibliographies. References to extensive sources are provided.

Springer Handbook of Global Navigation Satellite Systems

Peter J.G. Teunissen, Oliver Montenbruck (Eds.)

With 818 Figures and 193 Tables



Springer

Editors

Peter J.G. Teunissen
Curtin University
Department of Spatial Sciences
Perth, Australia
Perth, WA 6845, Australia

Oliver Montenbruck
German Aerospace Center (DLR)
Wessling, Germany
Münchener Str. 20
82234 Wessling, Germany

ISBN: 978-3-319-42926-7 e-ISBN: 978-3-319-42928-1
DOI 10.1007/978-3-319-42928-1
Library of Congress Control Number: 2017936757

© Springer International Publishing AG 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Production and typesetting: le-tex publishing services GmbH, Leipzig
Typography and layout: schreiberVIS, Seeheim
Illustrations: Hippmann GbR, Schwarzenbruck
Cover design: eStudio Calamar Steinen, Barcelona
Cover production: WMXDesign GmbH, Heidelberg
Printing and binding: Printer Trento s.r.l., Trento

Printed on acid free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Satellite navigation has become an integral part of our modern-day society and is used by people all over the world. Before 2000, practically only one system was fully operational and available, the American Global Positioning System (GPS). Its Russian counterpart GLONASS, built up during the times of the Cold War in the 1970s and 1980s, had problems with the lifetime of its satellites and decreased to five satellites in the late 1990s. When in 2002 Europe decided to develop its own global satellite system called Galileo, a global race began. Very soon, around 2020, four global satellite navigation systems will be available (GPS – USA, GLONASS – Russia, BeiDou – China, Galileo – Europe), as well as two regional systems (IRNSS/NavIC – India, QZSS – Japan) and more than six augmentation systems. Furthermore, there are probably more to come.

One may wonder whether we need so many systems, yet, who wants to stay on the sidelines in this high-tech area? Satellite navigation with its precise positioning, navigation, and timing (PNT) information is an enabling technology and an important factor in the economic impact of new applications. Precise satellite navigation time is used in the critical infrastructure (telecommunications, power supply, etc.) of many countries, and restricted and encrypted PNT services are a major element in governmental tasks and military applications.

A satellite navigation system is not comparable to other space missions. It does not have a lifetime of, say, 10 years, as is the case with other dedicated space missions that fulfil a specific goal for a limited scientific or technical part of our society. Satellite navigation systems need 10 to 20 years to be built up and are meant to last for many decades in the future. Moreover, they affect the life of each citizen – space for everybody!

In recent years, only a limited number of new textbooks addressing satellite navigation as a whole have become available. The reason is obvious; many systems were under construction, and the authors often had to restrict themselves to the general theory of satellite navigation, to the fundamentals, in order not to write a book that would already be outdated at the moment of publication. Separate books covered satellite navigation receivers, orbit determination, or navigation applications.

With the present book, we have the first *Handbook of Global Navigation Satellite Systems*, aiming to give the reader a full overview of satellite navigation, start-

ing from its fundamentals in the first part and covering in a total of seven parts the entire spectrum of satellite navigation knowledge and applications. Although some global satellite navigation systems are not yet fully completed, the Editors and authors were brave enough to include as much available information as possible.

More than 60 authors – all international experts and specialists in their field – have put together the latest state-of-the-art knowledge in the field in approximately 1400 pages. The names of the authors read like the *who is who* in satellite navigation. This is really an exhaustive reference work for one of the key technologies of science and engineering in the future and for those who want to know more about the background for their satellite navigation applications. It must already have been a huge amount of work for the Editors to find the right specialists, to convince them to contribute to such a handbook, to harmonize the different chapters in order to avoid duplications, and last but not least, to describe everything in a consistent way using the same nomenclature. This was a remarkable management exercise with an excellent outcome! Many figures and photographs illustrate the text, completing the handbook as *the* reference and resource in satellite navigation.

It is also excellent that the book is not only published as a hardcover version, but also as an eBook, which will enable the reader to do a quick search in this big exhaustive work. Each of the 41 chapters also contains a list of references from which the interested reader can start to dig even deeper into the topics. Nothing is missing. There are even two annexes that show the different receiver and data formats and outline the various GNSS parameters.

Each of the seven major parts of the book covers distinct aspects of Global Navigation Satellite Systems.

The reader confronted with satellite navigation for the first time can find a very useful, quick overview in Chap. 1. A very clear description and definition of the coordinate reference systems used follows. The chapter on clocks and the relativistic effects on GNSS is excellent; it covers everything from theory to practical



Guenter W. Hein
Former Head of ESA's
EGNOS and Galileo
Evolution Dept.
& Scientific Consultant of
The European Space
Agency

examples of clock data in space. Part A concludes with signal propagation in the atmosphere.

In Part B, the reader already finds a description of all the global, regional, and augmentation systems. The Editors have made great effort to reflect the latest status in the concluding developments of some of these systems.

Part C presents signal processing, receivers, and antennas, and the main effects: multipath and interferences. Even signal generators are described, a topic that is presently not well covered in the open literature.

In Part D, one can read all about the modern achievements in GNSS algorithms and, of course, about the LAMBDA method for the carrier-phase ambiguity resolution developed by one of the Editors in the past 15 years.

Part E entitled *Positioning and Navigation* shows the classical measurement modes of GNSS (absolute and differential positioning). All the new achievements of precise point positioning can be found here, as well the integration of GNSS with inertial navigation systems. The use of GNSS in aviation with its satellite-

based augmentation systems and attitude determination is described here. Here, you can also find a description of all GNSS applications on land and sea, as well as in aviation and space.

Part F is dedicated to surveying, geodesy and geodynamics, the first users of GNSS and the most demanding ones in terms of accuracy.

Finally, in Part G *GNSS Remote Sensing and Timing*, some of highly specialized GNSS applications are addressed: tomography, GNSS altimetry, as well as time and frequency transfer.

The list of acronyms and abbreviations and the glossary at the end of the book are very valuable.

There is no doubt that this handbook will quickly become *the* reference manual in satellite navigation – it contains all facets of this high-tech field. The Editors and the many authors have put together an exhaustive book – it is hard to find anything missing. I can only congratulate them for such a fine handbook!

Guenter W. Hein

Preface

In the mid-1970s, a team of inspired engineers gathered around Brad Parkinson to devise a novel navigation system, which finally became known as NAVSTAR, the Global Positioning System, or briefly GPS. Despite all creativity and visionary thought, the fathers of GPS could hardly imagine that this system would literally change the world. Originally conceived as a means for providing instantaneous positioning and timing to the United States' armed forces around the globe, it was soon realized that GPS would be equally beneficial for a wide range of civil navigation applications and could likewise serve as a system-of-opportunity for diverse types of scientific investigations. The Global Positioning System ultimately became a blueprint for a whole family of space-based navigation systems subsequently established by Russia, China, Europe, and Japan, which all build on the same key principles and technologies.

In response to its widespread use and overall impact, GPS has certainly found due attention and coverage in the popular, educational, and scholarly literature. Numerous textbooks and monographs have been published over the years, but more than 20 years have passed since GPS technology and usage was last summarized in a single comprehensive encyclopedia. Since then, substantial progress has been made and the world of global navigation satellite systems has changed dramatically. Obviously, the number of global navigation satellite systems (GNSS) has increased notably and so has the number of signals and services made available to their users. In parallel, the concepts of GNSS signal and data processing have continuously matured. This has both enabled new levels of performance (in terms of accuracy and timeliness) as well as a wide range of new application areas.

With the above background, we gratefully accepted the publisher's invitation (and challenge) to compile a dedicated *Handbook of Global Navigation Satellite Systems*, which presents a complete and rigorous overview of the fundamentals, methods, and applications of GNSS from today's perspective. After 4 years of work, this endeavor has come to an end, and we are proud to present the result of this work to our readers. With more than 40 chapters and roughly 1400 pages, the new handbook provides an exhaustive, single-source reference work and a state-of-the-art description of GNSS as a key technology for science and society at large. Key experts from a broad range of disciplines have contributed to cover the diverse aspects of GNSS and to summarize the respective

body-of-knowledge. This includes fundamental principles and technologies but likewise addresses the latest developments in the field. Throughout the handbook, due attention has been given to the new and emerging navigation satellite systems and the way they affect GNSS data processing and utilization.

Overall, the Handbook is structured in seven distinct parts, each comprising a set of four to eight individual chapters with a common range of topics.

Part A starts with a primer on global navigation satellite systems (GNSS), followed by chapters that cover the fundamentals of GNSS. Reference systems that form the backbone of positioning and navigation are discussed, as well as the essentials of GNSS satellite orbits and attitude. These topics are followed by a treatment of radio signals and modulations for GNSS, as well as a chapter on high-performance ground and space clocks that form another GNSS key technology. The part concludes with a description of the physics of atmospheric signal propagation and the changes experienced by GNSS signals when passing through the ionosphere and troposphere.

Part B gives a detailed description of the global and regional navigation satellite systems that are currently operational and/or under development. In addition to GPS, it covers the Russian Global'naya Navigatsionnaya Sputnikova Sistema (GLONASS), the European Galileo system, the Chinese BeiDou System (BDS), the Japanese Quasi-Zenith Satellite System (QZSS), and the Indian Regional Navigation Satellite System (IRNSS/NavIC). The architecture, the navigation signals, the space and control segments, and the services and performances of each system are described as well as their planned evolution. This part also covers the fundamentals and operations of satellite-based augmentation systems (SBASs).

Part C focuses on GNSS user equipment as well related aspects of signal multipath and interference. It discusses the basic architecture of hard and software receivers together with their digital signal processing principles. A dedicated chapter is devoted to GNSS antennas, including design options, performance aspects, and calibration. The multipath environment and its impact on code and phase measurements are described along with relevant mitigation techniques. Sources of GNSS signal interference are examined, and a systematic treatment of jamming and spoofing is provided together with a review of interference detection and mitigation techniques. The part is concluded with an

overview of the different GNSS simulators and a description of their key features.

Part D covers the fundamentals of the GNSS observation equations including the generic algorithms for GNSS parameter estimation and model validation. It starts with the basic observation equations for pseudo-range, carrier-phase, and Doppler measurements, and continues with a discussion of linear combinations and their applications. Then the undifferenced GNSS model is developed and used to provide an overview of the various absolute and relative positioning concepts. This is followed by a treatment of the fundamental GNSS estimation, filter, and ambiguity resolution algorithms, together with the corresponding batch and recursive methods for GNSS model validation.

Part E describes different methods of GNSS positioning and navigation together with their various applications. The single-receiver, precise point positioning concept is discussed first, highlighting its adjustment procedure as well as the required models needed to correct for systematic effects. Then the methods of differential positioning are described, including DGNSS, real-time kinematics (RTK) and network RTK. This is followed by a presentation of GNSS attitude determination methods and a discussion of GNSS integration with inertial measurement units. Subsequently, dedicated chapters provide an overview of GNSS applications in land, marine, air, and space environments, along with a discussion of ground-based augmentation systems (GBAS).

Part F describes how GNSS is used in surveying, geodesy, and geodynamics. It starts with an overview of the International GNSS Service (IGS) and a description of the various GNSS products that it offers. For surveying, this part describes how GNSS is used as a tool by the land, engineering, and hydrographic surveyor; for geodesy, it focuses on the role of GNSS in the Global Geodetic Observing System (GGOS), including GNSS-based reference frame implementation, Earth rotation, and sea level monitoring. For geodynamics, it describes the concepts and models used to relate active processes within the Earth to surface deformation observed with GNSS.

Part G, finally, covers GNSS remote sensing and timing. It describes how GNSS tropospheric sensing from ground and space can be used for short-term weather forecasting and long-term climate research. It also describes how GNSS ionospheric sensing contributes to space weather studies and how it helps to mitigate GNSS performance degradation. Furthermore, this part describes the principles of GNSS reflectometry together with methods to retrieve geophysical information from GNSS signals scattered or reflected at the Earth's surface. It concludes with a description of how

GNSS is used for accurate time and frequency dissemination and the comparison of distant clocks.

The main body of the book is complemented by an Annex that provides a detailed description of the most widely used GNSS data and product formats. It also offers a summary of relevant physical constants, key parameters of the GNSS constellations, and a compilation of the various GNSS signals. A Glossary covering GNSS-specific terminology is available at the end of the book to provide definitions of common terms that appear in the various chapters.

Overall, we are confident that the Handbook offers an invaluable source of knowledge for scientists, engineers, students, and institutions. It is likewise suited for readers who want to familiarize themselves with GNSS or one of its sub-disciplines and for experienced readers who aim at a deeper understanding of specific aspects. Unlike traditional textbooks, the individual chapters have been written by dedicated expert authors, who were selected based on their experience and background. Each chapter covers a specific aspect of GNSS in a largely self-contained manner and is thus well suited for standalone reading. However, individual chapters are still well connected through cross-references and follow a well-defined and transparent path for readers interested in studying all GNSS aspects in a step-by-step approach. Despite the overall size of the Handbook, space for each topic inevitably remains limited. Care has therefore been taken to complement each chapter with a thoroughly compiled list of bibliographic references covering both background literature and recent developments in the field. These may serve as a starting point for independent research and will enable readers to gain full insight into the numerous details of GNSS technology that cannot be addressed in a single-volume work.

At this stage, we would finally like to thank everyone who helped turn the vision of a new GNSS encyclopedia into reality. This includes, first of all, the more than 60 colleagues and authors who volunteered to contribute their expertise to this project. Despite a wealth of other duties, they all spent endless hours compiling the relevant information, preparing illustrating material, and adapting their work to the never-ending suggestions and change requests of the Editors. Their effort and patience have greatly contributed to achieving an up-to-date, concise, and consistent presentation of the whole world of GNSS in a single publication. We are also grateful to Ms Safoora Zaminpardaz of Curtin University, who assisted in the LaTeX conversion of numerous manuscripts and helped prepare various illustrations for the Handbook. Her assistance has taken substantial work off our shoulders and is thankfully acknowledged.

Our particular appreciation belongs to Ms J. Hinterberg and Ms J. Schwarz of Springer-Verlag, Heidelberg, as well as Ms A. Strohbach of le-tex, Leipzig, and her team for their excellent cooperation and continued support throughout all phases of this work. It was a great pleasure to work with them! Last but not least, we would both like to thank our families, who tolerated, without

major complaints, that we stole so many hours from them to work on this project. We are more than grateful for their patience and the continued backing received over the past years.

Peter J.G. Teunissen
Oliver Montenbruck

Perth, Australia
Munich, Germany

About the Editors

Peter J.G. Teunissen is Professor of Geodesy and Satellite Navigation at Curtin University (CU), Australia, and Delft University of Technology (TU Delft), The Netherlands. He obtained his Doctorate degree from TU Delft in 1985, following which he received a Constantijn and Christiaan Huygens Fellowship of The Netherlands Organization for the Advancement of Pure Research (NWO). In 1988, he became full Professor at TU Delft and held various academic positions including Vice-Dean of the Faculty of Civil Engineering and Geosciences, Head of the Aerospace Engineering Department, and Science Director of the Delft Institute of Earth Observation and Space Systems. He currently heads the Curtin University GNSS Research Centre, where his research is focused on developing theory, models, and algorithms for high-accuracy geospatial applications of new global and regional satellite navigation systems. He has authored numerous journal papers and textbooks in his field. His pioneering contributions include statistical and numerical methods of integer inference theory, innovative algorithms for multi-GNSS precise parameter estimation, and the early characterization and utilization of the Chinese BeiDou, the Indian IRNSS, and the Russian GLONASS CDMA system. His scientific contributions have been recognized through various awards, including the Bomford Prize, the Steven Hoogendijk Prize, and the Alexander von Humboldt Award. He serves on the Editorial Boards of several journals and is past Editor-in-Chief of the Journal of Geodesy. He has an Honorary Degree from the Chinese Academy of Sciences and is a Fellow of the International Association of Geodesy (IAG), the UK Royal Institute of Navigation (RIN), the US Institute of Navigation (ION), and the Royal Netherlands Academy of Sciences (KNAW).



Oliver Montenbruck is Head of the GNSS Technology and Navigation Group at DLR's German Space Operations Center, Oberpfaffenhofen. He studied Physics and Astronomy and received his diploma from Ludwig Maximilians University Munich in 1987. After joining DLR, he worked as a flight dynamics analyst for geostationary spacecraft as well as other near Earth and deep space missions. He received his PhD from the Technical University Munich in 1991. In 2004, he started teaching at the same university, where he received his habilitation (second doctorate) in 2006 and where he is presently engaged as an Associate Lecturer. His research activities comprise space borne GNSS receiver technology, autonomous navigation systems, spacecraft formation flying, and precise orbit determination. More recently, he has been focusing on the characterization of new satellite navigation systems and multi-GNSS processing. Pioneering contributions in this field included GIOVE and GPS signal investigations based on high gain antenna measurements, the build-up of the Cooperative Network for GNSS Observation (CONGO), the evaluation of triple-frequency signals, as well as the early characterization and utilization of the Chinese BeiDou navigation system. Oliver Montenbruck chairs the Multi-GNSS Working Group of the International GNSS Service (IGS) and coordinates the performance of the MGEX Multi-GNSS Project (MGEX). He has authored various textbooks and numerous technical papers related to his diverse fields of work. His scientific contributions have been recognized through various awards, including the DLR Senior Scientist Award, the Institute of Navigation's (ION) Tycho Brahe Award, and the GPS World Leadership Award.



List of Authors

Zuheir Altamimi

Institut National de l'Information Géographique
et Forestière (IGN)
Université Paris Diderot (LAREG)
Bâtiment Lamarck A et B, 35 rue Hélène Brion
75013 Paris, France
zuheir.altamimi@ign.fr

Felix Antreich

Federal University of Ceará (UFC)
Dept. of Teleinformatics Engineering
Campus do Pici – Bloco 725
CEP 60455-970, Fortaleza, Brazil
antreich@ieee.org

Ron Beard

US Naval Research Laboratory
Advanced Space PNT Branch
4555 Overlook Ave. SW
Washington, DC 20375, USA
ronald.beard@verizon.net

Alexey Bolkunov

Federal Space Agency (Roscosmos)
PNT Information and Analysis Center
4 Pionerskaya Street
141070 Korolyov, Russian Federation
alexei.bolkunov@glonass-iac.ru

Michael S. Braasch

Ohio University
School of Electrical Engineering & Computer
Science
1 Ohio University
Athens, OH 45701, USA
braaschm@ohio.edu

Thomas Burger

European Space Agency (ESA)
Galileo Project Office
Keplerlaan 1
2201 AZ, Noordwijk, The Netherlands
thomas.burger@esa.int

Estel Cardellach

Institute of Space Sciences
ICE (IEEC-CSIC)
Carrer de Can Magrans, S/N
08193 Cerdanyola del Valles, Spain
estel@ice.csic.es

James T. Curran

European Space Agency (ESA)
Keplerlaan 1
2201 AZ, Noordwijk, The Netherlands
jamestcurran@ieee.org

Pascale Defraigne

Royal Observatory of Belgium
Avenue Circulaire 3
1180 Brussels, Belgium
p.defraigne@oma.be

Bernd Eissfeller

Universität der Bundeswehr München
Inst. of Space Technology and Space Applications
Werner-Heisenberg-Weg 39
85577 Neubiberg, Germany
bernd.eissfeller@unibw.de

Gunnar Elgered

Chalmers University of Technology
Dept. of Earth and Space Sciences
Onsala Space Observatory, Observatorievägen 90
43992 Onsala, Sweden
gunnar.elgered@chalmers.se

Marco Falcone

European Space Agency (ESA)
Galileo Project Office
Keplerlaan 1
2201 AZ, Noordwijk, The Netherlands
marco.falcone@esa.int

Richard Farnworth

Eurocontrol Experimental Centre
Centre du Bois des Bordes – BP15
91222 Bretigny sur Orge, France
richard.farnworth@eurocontrol.int

Jay A. Farrell

University of California, Riverside
Dept. of Electrical and Computer Engineering
900 University Avenue
Riverside, CA 92521, USA
farrell@ece.ucr.edu

Jeff Freymueller

University of Alaska
Geophysical Institute
903 Koyukuk Drive
Fairbanks, AK 99775-7320, USA
jfreymueller@alaska.edu

A.S. Ganeshan

Indian Space Research Organization (ISRO)
ISRO Satellite Centre (ISAC)
Old Airport Road, Vimapura PIO
560017 Bangalore, India
asganeshan53@gmail.com

Steven Gao

University of Kent
School of Engineering and Digital Arts
Jennison Building
Canterbury, Kent, CT2 7NT, UK

Gabriele Giorgi

Technical University of Munich
Inst. for Communications and Navigation
Theresienstr. 90
80333 Munich, Germany
gabriele.giorgi@tum.de

Richard Gross

California Institute of Technology
Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, CA 91109, USA
richard.s.gross@jpl.nasa.gov

Jörg Hahn

European Space Agency (ESA)
Galileo Project Office
Keplerlaan 1
2201 AZ, Noordwijk, The Netherlands
joerg.hahn@esa.int

André Hauschild

German Aerospace Center (DLR)
German Space Operations Center
Münchener Str. 20
82234 Wessling, Germany
andre.hauschild@dlr.de

Grant Hausler

Geoscience Australia
Cnr Jerrabomberra Ave & Hindmarsh Drive
Symonston, ACT 2609, Australia
grant.hausler@ga.gov.au

Christopher J. Hegarty

The MITRE Corporation
202 Burlington Road
Bedford, MA 01886, USA
chegarty@mitre.org

Thomas Hobiger

Chalmers University of Technology
Onsala Space Observatory
Observatorievägen 90
43992 Onsala, Sweden
thomas.hobiger@chalmers.se

Urs Hugentobler

Technical University of Munich
Satellite Geodesy
Arcisstr. 21
80333 Munich, Germany
urs.hugentobler@bv.tum.de

Todd Humphreys

The University of Texas at Austin
Aerospace Engineering and Engineering
Mechanics, W.R. Woolrich Laboratories, C0600
210 East 24th Street
Austin, TX 78712-1221, USA
todd.humphreys@mail.utexas.edu

Norbert Jakowski

German Aerospace Center (DLR)
Institute of Communications and Navigation
Kalkhorstweg 53
17235 Neustrelitz, Germany
norbert.jakowski@dlr.de

Christopher Jekeli

Ohio State University
School of Earth Sciences
125 South Oval Mall
Columbus, OH 43210, USA
jekeli.1@osu.edu

Gary Johnston

Geoscience Australia
Cnr Jerrabomberra Ave & Hindmarsh Drive
Symonston, ACT 2609, Australia
gary.johnston@ga.gov.au

Allison Kealy

University of Melbourne
Dept. of Infrastructure Engineering
Grattan Street
Parkville, 3010 VIC, Australia
akealy@unimelb.edu.au

Satoshi Kogure

National Space Policy Secretariat, Cabinet Office
QZSS Strategy Office
1-6-1 Nagata-cho, Chiyoda-ku
100-8914 Tokyo, Japan
satoshi.kogure.e7f@cao.go.jp

Jan Kouba

Natural Resources Canada
Canadian Geodetic Survey
588 Booth Street
Ottawa, ON K1A 0Y7, Canada
kouba@rogers.com

François Lahaye

Natural Resources Canada
Canadian Geodetic Survey
588 Booth Street
Ottawa, ON K1A 0Y7, Canada
francois.lahaye@canada.ca

Richard B. Langley

University of New Brunswick
Dept. of Geodesy & Geomatics Engineering
15 Dineen Drive
Fredericton, NB E3B 5A3, Canada
lang@unb.ca

Ken MacLeod

Natural Resources Canada
Canadian Geodetic Survey
615 Booth Street
Ottawa, ON K1A 0E9, Canada
ken.macleod@canada.ca

Moazam Maqsood

Institute of Space Technology
Dept. of Electrical Engineering
1 Islamabad Highway
44000 Islamabad, Pakistan
moazam.maqsood@ist.edu.pk

Michael Meurer

German Aerospace Center (DLR)
Institute of Communications and Navigation
Münchener Str. 20
82234 Wessling, Germany
michael.meurer@dlr.de

Oliver Montenbruck

German Aerospace Center (DLR)
Münchener Str. 20
82234 Wessling, Germany
oliver.montenbruck@dlr.de

Terry Moore

University of Nottingham
Nottingham Geospatial Institute
Triumph Road
Nottingham, NG7 2TU, UK
terry.moore@nottingham.ac.uk

Dennis Odijk

Fugro Intersite B.V.
Dillenburgsingel 69
2263 HW, Leidschendam, The Netherlands
d.odijk@fugro.com

Thomas Pany

Universität der Bundeswehr München
Inst. of Space Technology and Space Applications
Werner-Heisenberg-Weg 39
85577 Neubiberg, Germany
thomas.pany@unibw.de

Mark G. Petovello

University of Calgary
Geomatics Engineering
2500 University Drive NW
Calgary, AB T2N 1N4, Canada
mark.petovello@ucalgary.ca

Sam Pullen

Stanford University
Dept. of Aeronautics and Astronautics
Durand Building, Room 250
Stanford, CA 94305-4035, USA
spullen@stanford.edu

Sergey Revnivkykh

RESHETNEV's Information Satellite Systems
Corporation
GLONASS Evolution Department
Mytischinskaya 3-16-60
129626 Moscow, Russian Federation
revnivkykh@iss-reshetnev.ru

Anna Riddell

Geoscience Australia
Cnr Jerrabomberra Ave & Hindmarsh Drive
Symonston, ACT 2609, Australia
anna.riddell@ga.gov.au

Antonio Rius

Institute of Space Sciences
ICE (IEEC-CSIC)
Carrer de Can Magrans, S/N
08193 Cerdanyola del Valles, Spain
rius@ice.csic.es

Chris Rizos

The University of New South Wales
School of Civil & Environmental Engineering
High Street
Kensington, NSW 2052, Australia
c.rizos@unsw.edu.au

Ken Senior

US Naval Research Laboratory
Advanced Space PNT Branch
4555 Overlook Ave. SW
Washington, DC 20375, USA
ken.senior@nrl.navy.mil

Alexander Serdyukov (deceased)**Tim Springer**

PosiTIm UG
In den Löser 15
64342 Seeheim-Jugenheim, Germany
tim.springer@positim.com

Peter Steigenberger

German Aerospace Center (DLR)
German Space Operations Center
Münchener Str. 20
82234 Wessling, Germany
peter.steigenberger@dlr.de

Jing Tang

China National Administration of GNSS and
Applications
17 Garden Road, Haidian District
100088 Beijing, China
blazingtangjing@163.com

Pierre Tétreault

Natural Resources Canada
Canadian Geodetic Survey
588 Booth Street
Ottawa, ON K1A 0Y7, Canada
pierre.tetreault@canada.ca

Peter J.G. Teunissen

Curtin University
Dept. of Spatial Sciences
Perth, WA 6845, Australia
p.teunissen@curtin.edu.au

Sandra Verhagen

Delft University of Technology
Faculty of Civil Engineering and Geosciences
Stevinweg 1
2628 CN, Delft, The Netherlands
a.a.verhagen@tudelft.nl

Todd Walter

Stanford University, GPS Lab
496 Lomita Mall Room 250
Stanford, CA 94305-4035, USA
twalter@stanford.edu

Lambert Wanninger

Technical University Dresden
Geodetic Institute
Helmholtzstr. 10
01069 Dresden, Germany
lambert.wanninger@tu-dresden.de

Jan P. Weiss

University Corporation for Atmospheric Research
COSMIC Program
3090 Center Green Drive
Boulder, CO 80301, USA
weissj@ucar.edu

Jan Wendel

Airbus DS GmbH
Navigation and Apps. Programmes
Robert-Koch-Str. 1
82024 Taufkirchen, Germany
jan.wendel@airbus.com

Jens Wickert

GFZ German Research Centre for Geosciences
Dept. of Geodesy
Telegrafenberg
14473 Potsdam, Germany
wickert@gfz-potsdam.de

Jong-Hoon Won

Inha University
Faculty of Electrical Engineering
100 Inharo, Nam-gu
Incheon, 22212, Korea
jh.won@inha.ac.kr

Yuanxi Yang

China National Administration of GNSS and
Applications
17 Garden Road, Haidian District
100088 Beijing, China
yuanxi_yang@163.com

Contents

List of Abbreviations	XXVII
------------------------------------	--------------

Part A Principles of GNSS

1 Introduction to GNSS	
<i>Richard B. Langley, Peter J.G. Teunissen, Oliver Montenbruck</i>	3
1.1 Early Satellite Navigation	3
1.2 Concept of GNSS Positioning	5
1.3 Modeling the Observations	10
1.4 Positioning Modes	13
1.5 Current and Developing GNSSs	16
1.6 GNSS for Science and Society at Large	19
References	22
2 Time and Reference Systems	
<i>Christopher Jekeli, Oliver Montenbruck</i>	25
2.1 Time	25
2.2 Spatial Reference Systems	31
2.3 Terrestrial Reference System	34
2.4 Celestial Reference System	44
2.5 Transformations Between ICRF and ITRF	46
2.6 Perspectives	55
References	56
3 Satellite Orbits and Attitude	
<i>Urs Hugentobler, Oliver Montenbruck</i>	59
3.1 Keplerian Motion	59
3.2 Orbit Perturbations	66
3.3 Broadcast Orbit Models	79
3.4 Attitude	85
References	87
4 Signals and Modulation	
<i>Michael Meurer, Felix Antreich</i>	91
4.1 Radiofrequency Signals	91
4.2 Spread Spectrum Technique and Pseudo Random Codes	97
4.3 Modulation Schemes	107
4.4 Signal Multiplexing	113
4.5 Navigation Data and Data-Free Channels	117
References	118
5 Clocks	
<i>Ron Beard, Ken Senior</i>	121
5.1 Frequency and Time Stability	122
5.2 Clock Technologies	127
5.3 Space-Qualified Atomic Standards	138
5.4 Relativistic Effects on Clocks	148

5.5	International Timescales	155
5.6	GNSS Timescales	158
	References	160
6	Atmospheric Signal Propagation	
	<i>Thomas Hobiger, Norbert Jakowski</i>	165
6.1	Electromagnetic Wave Propagation	165
6.2	Troposphere	168
6.3	Ionospheric Effects on GNSS Signal Propagation	177
	References	190
 Part B Satellite Navigation Systems		
7	The Global Positioning System (GPS)	
	<i>Christopher J. Hegarty</i>	197
7.1	Space Segment	197
7.2	Control Segment	203
7.3	Navigation Signals	205
7.4	Navigation Data and Algorithms	210
7.5	Time System and Geodesy	216
7.6	Services and Performance	216
	References	217
8	GLONASS	
	<i>Sergey Revnivkyh, Alexey Bolkunov, Alexander Serdyukov</i>	
	<i>Oliver Montenbruck</i>	219
8.1	Overview	219
8.2	Navigation Signals and Services	225
8.3	Satellites	232
8.4	Launch Vehicles	237
8.5	Ground Segment	238
8.6	GLONASS Open Service Performance	241
	References	243
9	Galileo	
	<i>Marco Falcone, Jörg Hahn, Thomas Burger</i>	247
9.1	Constellation	248
9.2	Signals and Services	250
9.3	Spacecraft	265
9.4	Ground Segment	269
9.5	Summary	270
	References	271
10	Chinese Navigation Satellite Systems	
	<i>Yuanxi Yang, Jing Tang, Oliver Montenbruck</i>	273
10.1	BeiDou Navigation Satellite Demonstration System (BDS-1)	275
10.2	BeiDou (Regional) Navigation Satellite System (BDS-2)	279
10.3	Performance of BDS-2	293
10.4	BeiDou (Global) Navigation Satellite System	297
10.5	Brief Introduction of CAPS	298
	References	301

11 Regional Systems	
<i>Satoshi Kogure, A.S. Ganeshan, Oliver Montenbruck</i>	305
11.1 Concept of Regional Navigation Satellite Systems	306
11.2 Quasi-Zenith Satellite System	306
11.3 Indian Regional Navigation Satellite System (IRNSS/NavIC)	321
References	334
12 Satellite Based Augmentation Systems	
<i>Todd Walter</i>	339
12.1 Aircraft Guidance	340
12.2 GPS Error Sources	343
12.3 SBAS Architecture	345
12.4 SBAS Integrity	349
12.5 SBAS User Algorithms	351
12.6 Operational and Planned SBAS Systems	353
12.7 Evolution of SBAS	358
References	360
 Part C GNSS Receivers and Antennas	
13 Receiver Architecture	
<i>Bernd Eissfeller, Jong-Hoon Won</i>	365
13.1 Background and History	366
13.2 Receiver Building Blocks	372
13.3 Multifrequency and Multisystem Receivers	391
13.4 Technology Trends	396
13.5 Receiver Types	397
References	399
14 Signal Processing	
<i>Jong-Hoon Won, Thomas Pany</i>	401
14.1 Overview and Scope	402
14.2 Received Signal Model	403
14.3 Signal Search and Acquisition	406
14.4 Signal Tracking	413
14.5 Time Synchronization and Data Demodulation	424
14.6 GNSS Measurements	428
14.7 Advanced Topics	434
References	440
15 Multipath	
<i>Michael S. Braasch</i>	443
15.1 The Impact of Multipath	444
15.2 Characterizing the Multipath Environment	444
15.3 Multipath Signal Models	448
15.4 Pseudorange and Carrier-Phase Error	450
15.5 Multipath Error Envelopes	450
15.6 Temporal Error Variation, Bias Characteristics and Fast Fading Considerations	453
15.7 Multipath Mitigation	455

15.8	Multipath Measurement	459
15.9	A Note About Multipath Impact on Doppler Measurements	466
15.10	Conclusions	466
	References	466
16	Interference	
	<i>Todd Humphreys</i>	469
16.1	Analysis Technique for Statistically Independent Interference	471
16.2	Canonical Interference Models	476
16.3	Quantization Effects	479
16.4	Specific Interference Waveforms and Sources	481
16.5	Spoofing.....	485
16.6	Interference Detection	491
16.7	Interference Mitigation	498
	References	501
17	Antennas	
	<i>Moazam Maqsood, Steven Gao, Oliver Montenbruck</i>	505
17.1	GNSS Antenna Characteristics.....	506
17.2	Basic GNSS Antenna Types	509
17.3	Application-Specific GNSS Antennas.....	513
17.4	Multipath Mitigation	519
17.5	Antennas for GNSS Satellites.....	523
17.6	Antenna Measurement and Calibration	527
	References	531
18	Simulators and Test Equipment	
	<i>Mark G. Petovello, James T. Curran</i>	535
18.1	Background	537
18.2	RF-Level Simulators	543
18.3	IF-Level Simulators.....	546
18.4	Record and Playback Systems	549
18.5	Measurement-Level Simulators	552
18.6	Combining Live and Simulated Data.....	554
18.7	Other Considerations	556
18.8	Summary	557
	References	557

Part D GNSS Algorithms and Models

19	Basic Observation Equations	
	<i>André Hauschild</i>	561
19.1	Observation Equations	561
19.2	Relativistic Effects	564
19.3	Atmospheric Signal Delays.....	565
19.4	Carrier-Phase Wind-Up	569
19.5	Antenna Phase-Center Offset and Variations	572
19.6	Signal Biases	576
19.7	Receiver Noise and Multipath	578
	References	579

20 Combinations of Observations	
<i>André Hauschild</i>	583
20.1 Fundamental Equations	583
20.2 Combinations of Single-Satellite and Single-Receiver Observations	586
20.3 Combinations of Multisatellite and Multireceiver Observations	594
20.4 Pseudorange Filtering	601
References	603
21 Positioning Model	
<i>Dennis Odijk</i>	605
21.1 Nonlinear Observation Equations	606
21.2 Linearization of the Observation Equations	609
21.3 Point Positioning Models	612
21.4 Relative Positioning Models	623
21.5 Differenced Positioning Models	631
21.6 The Positioning Concepts Related	633
References	635
22 Least-Squares Estimation and Kalman Filtering	
<i>Sandra Verhagen, Peter J.G. Teunissen</i>	639
22.1 Linear Least-Squares Estimation	639
22.2 Optimal Estimation	641
22.3 Special Forms of Least Squares	644
22.4 Prediction and Filtering	650
22.5 Kalman Filtering	653
References	659
23 Carrier Phase Integer Ambiguity Resolution	
<i>Peter J.G. Teunissen</i>	661
23.1 GNSS Ambiguity Resolution	662
23.2 Rounding and Bootstrapping	666
23.3 Linear Combinations	669
23.4 Integer Least-Squares	673
23.5 Partial Ambiguity Resolution	677
23.6 When to Accept the Integer Solution?	678
References	683
24 Batch and Recursive Model Validation	
<i>Peter J.G. Teunissen</i>	687
24.1 Modeling and Validation	687
24.2 Batch Model Validation	689
24.3 Testing for a Bias	692
24.4 Testing Procedure	705
24.5 Recursive Model Validation	710
References	717
 Part E Positioning and Navigation	
25 Precise Point Positioning	
<i>Jan Kouba, François Lahaye, Pierre Tétreault</i>	723
25.1 PPP Concept	724

25.2	Precise Positioning Correction Models	726
25.3	Specific Processing Aspects	735
25.4	Implementations	741
25.5	Examples	743
25.6	Discussion	746
	References	747
26	Differential Positioning	
	<i>Dennis Odijk, Lambert Wanninger</i>	753
26.1	Differential GNSS: Concepts	753
26.2	Differential Navigation Services	760
26.3	Real-Time Kinematic Positioning	763
26.4	Network RTK	774
	References	778
27	Attitude Determination	
	<i>Gabriele Giorgi</i>	781
27.1	Six Degrees of Freedom	781
27.2	Attitude Parameterization	784
27.3	Attitude Estimation from Baseline Observations	787
27.4	The GNSS Attitude Model	790
27.5	Applications	798
27.6	An Overview of GNSS/INS Sensor Fusion for Attitude Determination	804
	References	806
28	GNSS/INS Integration	
	<i>Jay A. Farrell, Jan Wendel</i>	811
28.1	State Estimation Objectives	812
28.2	Inertial Navigation	813
28.3	Inertial Sensors	815
28.4	Strapdown Inertial Navigation	818
28.5	Analysis of Error Effects	822
28.6	Aided Navigation	824
28.7	State Estimation	824
28.8	GNSS and Aided INS	825
28.9	Detailed Example	828
28.10	Alternative Estimation Methods	835
28.11	Looking Forward	838
	References	839
29	Land and Maritime Applications	
	<i>Allison Kealy, Terry Moore</i>	841
29.1	Land-Based Applications of GNSS	842
29.2	Rail Applications	856
29.3	Maritime Applications	863
29.4	Outlook	873
	References	873
30	Aviation Applications	
	<i>Richard Farnworth</i>	877
30.1	Overview	878

30.2	Standardising GNSS for Aviation	881
30.3	Evolution of the Flight Deck	884
30.4	From the RNP Concept to PBN	886
30.5	GNSS Performance Requirements	888
30.6	Linking the PBN Requirements and the GNSS Requirements	891
30.7	Flight Planning and NOTAMs	897
30.8	Regulation and Certification	897
30.9	Military Aviation Applications	898
30.10	Other Aviation Applications of GNSS	899
30.11	Future Evolution	900
	References	901
31	Ground Based Augmentation Systems	
	<i>Sam Pullen</i>	905
31.1	Components	906
31.2	An Overview of Local Area Approaches	907
31.3	Ground-Based Augmentation Systems	909
31.4	Augmentation via Ranging Signals Pseudolites	928
31.5	Outlook	930
	References	930
32	Space Applications	
	<i>Oliver Montenbruck</i>	933
32.1	Flying High	933
32.2	Spacecraft Navigation	938
32.3	Formation Flying and Rendezvous	951
32.4	Other Applications	957
	References	959
 Part F Surveying, Geodesy and Geodynamics		
33	The International GNSS Service	
	<i>Gary Johnston, Anna Riddell, Grant Hausler</i>	967
33.1	Mission and Organization	967
33.2	Components	969
33.3	IGS Products	972
33.4	Pilot Projects and Experiments	976
33.5	Outlook	981
	References	981
34	Orbit and Clock Product Generation	
	<i>Jan P. Weiss, Peter Steigenberger, Tim Springer</i>	983
34.1	Global Tracking Network	984
34.2	Models	985
34.3	POD Process	992
34.4	Estimation Strategies	993
34.5	Software	1000
34.6	Products	1001
34.7	Outlook	1005
	References	1006

35 Surveying	
<i>Chris Rizos</i>	1011
35.1 Precise Positioning Techniques.....	1013
35.2 Geodetic and Land Surveying.....	1023
35.3 Engineering Surveying.....	1029
35.4 Hydrographic Surveying.....	1033
References	1035
36 Geodesy	
<i>Zuheir Altamimi, Richard Gross</i>	1039
36.1 GNSS and IAG's Global Geodetic Observing System.....	1039
36.2 Global and Regional Reference Frames.....	1044
36.3 Earth Rotation, Polar Motion, and Nutation.....	1054
References	1059
37 Geodynamics	
<i>Jeff Freymueller</i>	1063
37.1 GNSS for Geodynamics.....	1064
37.2 History and Establishment of GNSS Networks for Geodynamics.....	1067
37.3 Rigid Plate Motions.....	1071
37.4 Plate Boundary Deformation and the Earthquake Cycle.....	1073
37.5 Seismology.....	1078
37.6 Volcano Deformation.....	1088
37.7 Surface Loading Deformation.....	1091
37.8 The Multi-GNSS Future.....	1099
References	1100
 Part G GNSS Remote Sensing and Timing	
38 Monitoring of the Neutral Atmosphere	
<i>Gunnar Elgered, Jens Wickert</i>	1109
38.1 Ground-Based Monitoring of the Neutral Atmosphere.....	1110
38.2 GNSS Radio Occultation Measurements.....	1120
38.3 Outlook.....	1132
References	1133
39 Ionosphere Monitoring	
<i>Norbert Jakowski</i>	1139
39.1 Ground-Based GNSS Monitoring.....	1140
39.2 Space-Based GNSS Monitoring.....	1144
39.3 GNSS-Based 3-D-Tomography.....	1147
39.4 Scintillation Monitoring.....	1148
39.5 Space Weather.....	1152
39.6 Coupling with Lower Geospheres.....	1156
39.7 Information and Data Services.....	1159
References	1159
40 Reflectometry	
<i>Antonio Rius, Estel Cardellach</i>	1163
40.1 Receivers.....	1164
40.2 Models.....	1167

40.3	Applications.....	1172
40.4	Spaceborne Missions	1182
	References	1183
41	GNSS Time and Frequency Transfer	
	<i>Pascale Defraigne</i>	1187
41.1	GNSS Time and Frequency Dissemination	1187
41.2	Remote Clock Comparisons	1191
41.3	Hardware Architecture and Calibration	1197
41.4	Multi-GNSS Time Transfer	1201
41.5	Conclusions	1203
	References	1204
	Annex A: Data Formats	1207
	Annex B: GNSS Parameters	1233
	About the Authors	1241
	Detailed Contents	1251
	Glossary of Defining Terms	1275
	Subject Index	1303

List of Abbreviations

A

A-GNSS	assisted GNSS
A-PNT	alternative positioning navigation and timing
ABAS	aircraft based augmentation system
AC	analysis center
ACES	atom clock ensemble in space
ADC	analog-to-digital converter
ADEV	Allan deviation
ADF	automatic direction finding
ADOP	ambiguity dilution of precision
ADS	automatic dependent surveillance
AEP	architecture evolution plan
AFS	atomic frequency standard
AFSCN	air force satellite control network
AGC	automatic gain control
AGGA	advanced GPS/GLONASS ASIC
AIUB	Astronomical Institute of the University of Bern
AKM	apogee kick motor
AltBOC	alternative BOC
AM	amplitude modulation
ANTEX	antenna exchange (format)
AOCS	attitude and orbit control system
APL	airport pseudolite
APV	approach with vertical guidance
ARP	antenna reference point
ARW	angular random walk
ASIC	application specific integrated circuit
ATV	automated transfer vehicle
AUT	antenna under test
AWGN	additive white Gaussian noise

B

BAW	bulk acoustic wave
BC	Barker code
BCH	Bose–Chaudhuri–Hocquenghem (code)
BCRS	barycentric celestial reference system
BDS	BeiDou Navigation Satellite System
BDT	BeiDou time
BGD	broadcast group delay
BIH	Bureau International de l'Heure
BIPM	Bureau International des Poids et Mesures
BLUE	best linear unbiased estimation
BLUP	best linear unbiased prediction
BNR	bias-to-noise ratio
BOC	binary offset carrier
BPSK	binary phase-shift keying

C

CAPS	Chinese Area Positioning System
CASM	coherent adaptive sub-carrier modulation
CBOC	composite binary offset carrier
CCD	code-carrier divergence
CCIR	Comité Consultatif International des Radiocommunications
CDGNSS	carrier-phase differential GNSS
CDMA	code division multiple access
CEP	circular error probable
CFIT	controlled flight into terrain
CGCS	China Geodetic Coordinate System
CIO	celestial intermediate origin
CL	long code
CM	moderate-length code
CMC	code-minus-carrier
CMCU	clock monitoring and comparison unit
CMOS	complementary metal oxide semiconductor
CMS	constrained maximum success-rate
CNAV	civil navigation message
CODE	Center for Orbit Determination in Europe
COG	center-of-gravity
CoM	center-of-mass
CoN	center-of-network
CONUS	conterminous United States
COO	cell-of-origin
CORS	continuously operating reference station
COSPAS	Cosmicheskaya Sistema Poiska Avaryinyh Sudov (space system for search of distress vessels and airplanes)
COTS	commercial-off-the-shelf
CPT	coherent population trapping
CPU	central processing unit
CRC	cyclic redundancy check
CRF	celestial reference frame
CRPA	controlled radiation pattern antenna
CRS	celestial reference system
CS	Commercial Service
CS	control segment
CSAC	chip scale atomic clock
CSK	code shift keying
CTP	conventional terrestrial pole

D

DAB	digital audio broadcast
DC	data center
DCB	differential code bias
DCFBS	digital cesium beam frequency standard
DD	double-difference

DDM	delay-Doppler-map
DEM	digital elevation model
DGNSS	differential GNSS
DH	decision height
DIODE	DORIS immediate orbit on-board determination
DLL	delay lock loop
DLR	Deutsches Zentrum für Luft- und Raumfahrt
DMA	Defense Mapping Agency
DME	distance measuring equipment
DOP	dilution of precision
DORIS	Doppler orbitography and radiopositioning integrated by satellite
DQM	data quality monitoring
DSP	digital signal processor
DVB	digital video broadcasting

E

EAL	Echelle atomique libre (free atomic scale)
ECEF	Earth-centered Earth-fixed
ECI	Earth-centered inertial
ECMWF	European Centre for Medium-Range Weather Forecasts
ECOM	Empirical CODE Orbit Model
EELV	evolved expendable launch vehicles
EGNOS	European Geostationary Navigation Overlay Service
EIRP	effective isotropic radiated power
EKF	extended Kalman filter
EO	Earth observation
EPB	equatorial plasma bubble
ESA	European Space Agency
EUV	extreme ultraviolet

F

FAA	US Federal Aviation Administration
FAR	full ambiguity resolution
FBR	front-to-back ratio
FCC	Federal Communications Commission
FDMA	frequency division multiple access
FE	front end
FEC	forward error correction
FFT	fast Fourier transform
FK5	Fundamental Katalog 5
FLDR	flicker drift
FLFR	flicker frequency (noise)
FLL	frequency lock loop
FLPH	flicker phase (noise)
FMS	flight management system
FNBW	first-null beam width
FOC	full operational capability
FOG	fiber optic gyroscope
FPGA	field programmable gate array
FRPA	fixed radiation pattern antenna
FTE	flight technical error

G

GAGAN	GPS-aided GEO Augmented Navigation
GAST	Greenwich apparent sidereal time
GBAS	ground-based augmentation system
GCC	Galileo Control Centre
GCRS	Geocentric Celestial Reference System
GDGPS	global differential GPS
GEO	geostationary Earth orbit
GFZ	Deutsches GeoForschungsZentrum
GGTO	GPS-to-Galileo time offset
GIM	global ionospheric map
GIOVE	Galileo In-Orbit Validation Element
GIS	geographic information system
GIVE	grid ionospheric vertical error
GLONASS	Global'naya Navigatsionnaya Sputnikova Sistema (Russian Global Navigation Satellite System)
GLST	GLONASS System Time
GMS	ground mission segment
GNSS	global navigation satellite system
GPS	Global Positioning System
GPST	GPS Time
GPT	global pressure and temperature (model)
GRAM	GPS receiver application module
GRAS	ground-based regional augmentation system
GRAS	GNSS receiver for atmospheric sounding
GST	Galileo System Time
GTRS	Geocentric Terrestrial Reference System

H

HDOP	horizontal dilution of precision
HEO	highly elliptical orbit
HOW	hand-over word
HPBW	half-power beam width

I

I/Q	in-phase/quadrature
IAU	International Astronomical Union
IB	integer bootstrapping
IBLS	integrity beacon landing system
ICAO	International Civil Aviation Organization
ICD	interface control document
ICRF	International Celestial Reference Frame
ICRS	International Celestial Reference System
IEEE	Institute of Electrical and Electronics Engineers
IERS	International Earth Rotation and Reference Systems Service
IF	intermediate frequency
IFA	inverted-F antenna
IGP	ionospheric grid point
IGS	International GNSS Service
IGSO	inclined geo-synchronous orbit
IIP	instantaneous impact point
ILS	integer least-squares

ILS	instrument landing system
ILS	International Latitude Service
IMU	inertial measurement unit
INS	inertial navigation system
InSAR	interferometric synthetic aperture radar
IOD	issue-of-data
IODC	issue-of-data clock
IODE	issue-of-data ephemeris
IONEX	ionosphere exchange (format)
IOP	intensity optical pumping
IOT	in-orbit test
IOV	in-orbit validation
IPP	ionospheric pierce point
IR	integer rounding
IRI	international reference ionosphere
IRM	IERS reference meridian
IRNSS	Indian Regional Navigation Satellite System
IRP	international reference pole
ISB	intersystem bias
ISC	intersignal correction
ISS	International Space Station
ITRF	International Terrestrial Reference Frame
ITRS	International Terrestrial Reference System
ITS	intelligent transport system
ITU	International Telecommunication Union
IUGG	International Union of Geodesy and Geophysics

J

JAXA	Japan Aerospace Exploration Agency
JD	Julian day/date
JPL	Jet Propulsion Laboratory

K

KASS	Korean Augmentation Satellite System
KF	Kalman filter

L

L-AII	Legacy Accuracy Improvement Initiative
LADGNSS	local area differential GNSS
LAMBDA	least-squares ambiguity decorrelation adjustment
LEO	low Earth orbit
LEOP	launch and early orbit phase
LHCP	left-hand circular polarized
LIDAR	light detection and ranging
LNA	low-noise amplifier
LNAV	legacy navigation message
LOS	line-of-sight
LQG	linear quadratic Gaussian
LRA	laser retro-reflector array
LTT	laser time transfer

M

MAP	maximum a posteriori
MC	master clock
MCS	master control station
MDB	minimal detectable bias
MEMS	micro-electromechanical system
MEO	medium Earth orbit
MJD	modified Julian day/date
MLE	maximum likelihood estimation
MLS	microwave landing system
MMP	minimum mean penalty
MMS	Magnetosphere Multiscale Mission
MOT	magneto-optical trap
MPR	multipath rejection ratio
MQM	measurements quality monitoring
MS	monitoring station
MSAS	Multi-Function Satellite Augmentation System
MSS	mean squared slope

N

NAD	North American Datum
NANU	notice advisory to NAVSTAR users
NAQU	notice advisory to QZSS users
NASA	National Aeronautics and Space Administration
NCO	numerically controlled oscillator
NCP	North celestial pole
NDB	nondirectional beacon
NEP	North ecliptic pole
NGA	National Geospatial-Intelligence Agency
NH	Neuman-Hofman (code)
NIST	National Institute of Standards and Technology
NMCT	navigation message correction table
NMEA	National Marine Electronics Association
NMF	Niell mapping function
NOTAM	notice to airmen
NPA	nonprecision approach
NRL	Naval Research Lab
NSE	navigation system error
NUDET	nuclear detection (payload)
NWM	numerical weather model
NWP	numerical weather prediction

O

OCS	operational control system
OCX	next generation operational control segment of GPS
OCXO	oven controlled crystal oscillator
OEM	original equipment manufacturer
OS	Open Service
OSPF	orbitography and synchronization processing facility
OWCP	one-way carrier-phase technique

P

PAR	partial ambiguity resolution
PBN	performance based navigation
PBO	plate boundary observatory
PCB	printed circuit board
PCO	phase center offset
PCV	phase center variation
PDA	personal digital assistant
PDF	probability density function
PDOP	position dilution of precision
PF	particle filter
PHM	passive hydrogen maser
PLL	phase lock loop
PM	phase modulation
PMF	probability mass function
PNT	positioning, navigation and timing
POD	precise orbit determination
PPD	personal privacy device
PPP	precise point positioning
PPS	precise positioning service
PPS	pulse per second
PRC	pseudorange correction
PRN	pseudo-random noise
PSD	power spectral density
PVT	position, velocity and time

Q

QHA	quadrifilar helix antenna
QPSK	quadrature phase-shift keying
QZSS	Quasi-Zenith Satellite System

R

RAAN	right ascension of ascending node
RAFS	rubidium atomic frequency standard
RAIM	receiver autonomous integrity monitoring
RDSS	radio determination satellite service
RF	radio frequency
RFI	radio frequency interference
RFSA	Russian Federal Space Agency
RHCP	right-hand circular polarized
RINEX	receiver independent exchange (format)
RLG	ring laser gyroscope
RMS	root mean square
RNAV	area navigation
RNP	required navigation performance
RNSS	radio navigation satellite service
RNSS	regional navigation satellite system
RO	radio occultation
RRC	range-rate correction
RSS	root-sum-square
RTAC	Real-Time Analysis Center
RTCA	Radio Technical Commission for Aeronautics
RTCM	Radio Technical Commission for Maritime Services

RTI	Rayleigh-Taylor instability
RTK	real-time kinematic
RTS	real-time service
RWDR	random walk drift
RWFR	random walk frequency (noise)
RWPH	random walk phase (noise)

S

SA	selective availability
SAASM	selective availability anti-spoofing module
SAR	synthetic aperture radar
SAR	search and rescue
SARPS	standards and recommended practices
SAW	surface acoustic wave
SBAS	satellite-based augmentation system
SD	single-difference
SDA	strapdown algorithm
SDCM	System for Differential Corrections and Monitoring
SDM	signal deformation monitoring
SDR	software defined radio
SEL	single event latch-up
SEU	single event update
SIGI	Space Integrated GPS/Inertial navigation system
SINEX	solution independent exchange (format)
SISO	single-input-single-output
SISRAD	signal-in-space receive and decode
SISRE	signal-in-space range error
SLAM	simultaneous location and mapping
SLR	satellite laser ranging
SLTA	straight line tangent point altitude
SNR	signal-to-noise ratio
SoC	system-on-a-chip
SOFA	standards of fundamental astronomy
SP3	Standard Product 3 (format)
SPAD	single photon avalanche diode
SPP	single point positioning
SPS	standard positioning service
SRP	solar radiation pressure
ST	system time
STEC	slant total electron content
SV	space vehicle
SVN	space vehicle number

T

TACAN	tactical air navigation (system)
TAI	International Atomic Time
TASS	TDRSS augmentation service for satellites
TCB	barycentric coordinate time
TCG	Geocentric Coordinate Time
TCXO	temperature compensated crystal oscillator
TDB	barycentric dynamic time

TDRSS	tracking and data relay satellite system
TDT	terrestrial dynamic time
TEC	total electron content
TGD	timing group delay
TID	total ionization dose
TID	traveling ionospheric disturbance
TIO	terrestrial intermediate origin
TKS	time keeping system
TLM	telemetry (word)
TMBOC	time multiplexed binary offset carrier
TOA	time-of-arrival
TRF	terrestrial reference frame
TT	terrestrial time
TTA	time-to-alert
TTFF	time-to-first-fix
TWSTFT	two-way satellite time and frequency transfer
TWTA	traveling wave tube amplifier

U

UAV	unmanned aerial vehicle
UDRE	user differential range error
UERE	user equivalent range error
UHF	ultra-high frequency
UMPI	uniformly most powerful invariant
UNAVCO	University NAVSTAR Consortium
UNB	University of New Brunswick
URSI	International Union of Radio Science
USGS	United States Geological Survey
USNO	United States Naval Observatory
UT	Universal Time

UTC	Coordinated Universal Time
UWB	ultra-wideband

V

VDB	VHF data broadcast
VDOP	vertical dilution of precision
VHF	very high frequency
VLBI	very long baseline interferometry
VMF	Vienna mapping function
VNA	vector network analyzer
VOR	VHF omnidirectional range
VPL	vertical protection level
VRE	vibration rectification error
VRW	velocity random walk
VSWR	voltage standing wave ratio
VTEC	vertical total electron content

W

WAAS	Wide Area Augmentation System
WAGE	wide area GPS enhancement
WGS	World Geodetic System
WHPH	white phase (noise)
WLS	weighted least-squares

Z

ZHD	zenith hydrostatic delay
ZTD	zenith troposphere delay
ZWD	zenith wet delay

Principle

Part A

Part A Principles of GNSS

1 Introduction to GNSS

Richard B. Langley, Fredericton, Canada
Peter J.G. Teunissen, Perth, Australia
Oliver Montenbruck, Wessling, Germany

2 Time and Reference Systems

Christopher Jekeli, Columbus, USA
Oliver Montenbruck, Wessling, Germany

3 Satellite Orbits and Attitude

Urs Hugentobler, Munich, Germany
Oliver Montenbruck, Wessling, Germany

4 Signals and Modulation

Michael Meurer, Wessling, Germany
Felix Antreich, Fortaleza, Brazil

5 Clocks

Ron Beard, Washington, USA
Ken Senior, Washington, USA

6 Atmospheric Signal Propagation

Thomas Hobiger, Onsala, Sweden
Norbert Jakowski, Neustrelitz, Germany

Introduction

1. Introduction to GNSS

Richard B. Langley, Peter J.G. Teunissen, Oliver Montenbruck

This chapter is a primer on global navigation satellite systems (GNSSs). It assumes no prior knowledge of the systems or how they work. All of the key concepts of satellite-based positioning, navigation, and timing (PNT) are introduced with pointers to subsequent chapters for further details. The chapter begins with a history of PNT using satellites and then introduces the concept of positioning using measured ranges between a receiver and satellites. The basic observation equations are then described along with the associated error budgets. Subsequently, the various GNSSs now in operation and in development are briefly overviewed. The chapter concludes with a discussion of the relevance and importance of GNSS for science and society at large.

1.1	Early Satellite Navigation	3
1.2	Concept of GNSS Positioning	5
1.2.1	Ranging Measurements	5
1.2.2	Range-Based Positioning	6
1.2.3	Pseudorange Positioning	7
1.2.4	Precision of Position Solutions	8
1.2.5	GNSS Observation Equations	10
1.3	Modeling the Observations	10
1.3.1	Satellite Orbit and Clock Information	10
1.3.2	Atmospheric Propagation Delay	11
1.4	Positioning Modes	13
1.4.1	Precise Point Positioning	13
1.4.2	Code Differential Positioning	14
1.4.3	Differential Carrier Phase	14
1.5	Current and Developing GNSSs	16
1.5.1	Global Navigation Satellite Systems	16
1.5.2	Regional Navigation Satellite Systems	18
1.5.3	Satellite-Based Augmentation Systems	19
1.6	GNSS for Science and Society at Large	19
	References	22

1.1 Early Satellite Navigation

We will introduce the basic concepts of the Navstar Global Positioning System (GPS) and the other global navigation satellite systems (GNSSs) in operation and under development but it will be helpful if we first view them in a historical perspective. Determining the positions of points on the Earth's surface using observations of distant objects has been carried out for hundreds of years. Reflecting mirrors on mountaintops gave way to using high-altitude flares and rockets. And, of course, celestial navigation using observations of the Sun, stars, and planets has been used for centuries. However, it was only with the dawning of the space age that it became possible to develop a global system for high accuracy positioning and navigation.

Sometimes we refer to these systems as space-based systems. They can be broadly classified into optical techniques and radio techniques. Both kinds of system were pioneered in the late 1950s and 1960s.

Optical techniques are those techniques that utilize the visible part of the electromagnetic spectrum and in addition to astronomical positioning using a theodolite or sextant, include ground-based imaging of orbiting satellites and satellite laser ranging (SLR). Although still an important source of information for satellite orbit determination and surveillance, imaging of satellites against background stars for geodetic positioning has been superseded by other techniques. However, SLR still plays a prominent role in geodetic positioning and celestial mechanics.

Several radio systems were developed for satellite tracking and orbit determination. In the United States, these systems included radar, the Goddard Space Flight Center Range and Range Rate (GRARR) system, and NASA's Minitrack system [1.1]. In addition to their role in orbit determination, the systems were utilized for tracking camera calibration and directly for geode-

tic positioning. The US Army's Sequential Collation of Range (SECOR) system, on the other hand, was developed specifically for positioning purposes.

But the most successful satellite-based positioning system and one that overlapped with the development of GPS, was Transit [1.2, 3]. Also known as the US Navy Navigation Satellite System, Transit was the world's first satellite-based positioning system to operate globally. The system evolved from the efforts to track the Soviet Union's Sputnik I, the first artificial Earth-orbiting satellite. By measuring the Doppler frequency shift of the 20 MHz radio signals received from the satellite at a known location, the orbit of the satellite could be worked out. And shortly thereafter, researchers determined that if the orbit of a satellite was known, then the position of a receiver could be determined from the shift. That realization led to the development of Transit, with the first experimental satellite being launched in 1959. Initially classified, the system was made available to civilians in 1967 and was widely used for navigation and precise positioning until it was shut down in 1996. The Soviet Union developed a similar system called Tsikada and a special military version called Parus [1.4, 5]. These systems are also assumed to be no longer in use – at least for navigation.

A series of Transit prototype and research satellites was launched between 1959 and 1964 with the first fully operational satellite, Transit 5-BN-2, launched on 5 December 1963. The first Oscar-class Transit satellite (NNS O-1, Fig. 1.1), was brought into orbit on 6

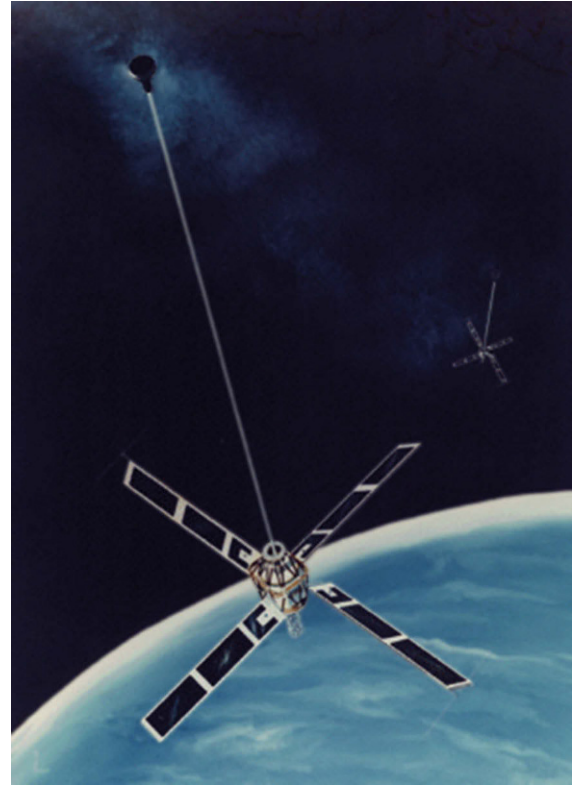


Fig. 1.1 US Navy Transit navigation satellite of the *Oscar* series (named after the phonetic code word for the letter *O*, or *operational*) (courtesy of US Navy)

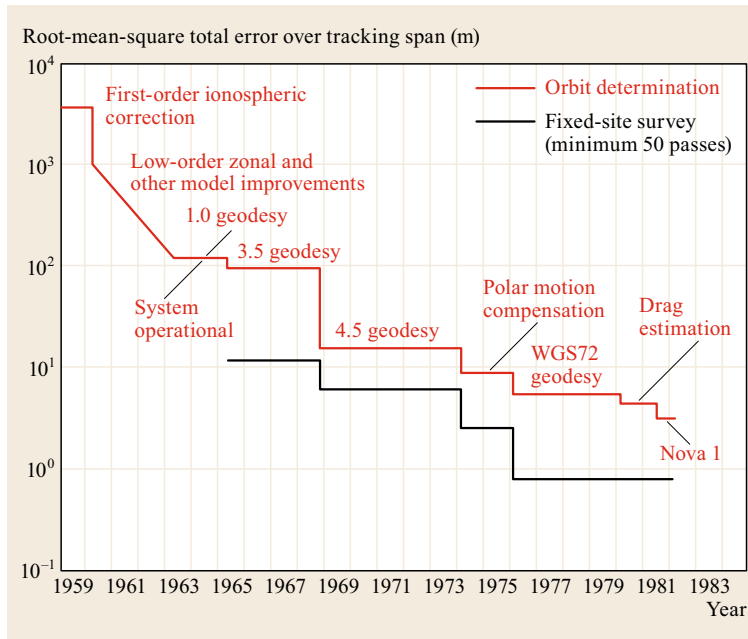


Fig. 1.2 Accuracy improvements over time with US Navy Transit satellites (after [1.3])

October 1964 and 24 operational satellites were subsequently launched. The last pair of Transit satellites, NNS O-25 and O-31, was launched on 25 August 1988.

Transit navigation required the measurements of the satellite signal's Doppler shift for a complete pass that could take up to about 18 min from horizon to horizon. At the conclusion of the pass, the latitude and longitude of the receiver, the position fix, could be determined. With five operational satellites, the mean time between fixes at a mid-latitude site was around 1 h. Eventually, as the orbits of the satellites became better determined, two-dimensional (2-D) position fix accuracies of several tens of meters were possible from a single satellite pass. By recording data from a number of passes over a few days from a fixed site on land, three-dimensional (3-D) accuracies better than one meter were possible and Doppler-based control points for mapping were established in many countries and the Canadian north, in particular, saw significant use of Transit for geodetic purposes.

1.2 Concept of GNSS Positioning

1.2.1 Ranging Measurements

GNSS signals are electromagnetic waves propagating at the speed of light. Signal frequencies in the radio spectrum between about 1.2 and 1.6 GHz (a part of the so-called L-band) have been selected for these signals since these enable measurements of adequate precision, allow for reasonably simple user equipment and do not suffer from attenuation in the atmosphere under common weather conditions. At the given frequencies, GNSS signals have a wavelength of about 19–25 cm. Similar to early radio navigation systems such as Transit, GNSSs provide signals on at least two different frequencies for compensation of ionospheric delays in their measurements.

A distinct feature of all GNSS signals is the modulation of the harmonic radio wave (termed the *carrier*) with a characteristic pseudorandom noise (PRN) code. This code is essentially a binary sequence of zeros and ones with no obvious pattern or regularity. The sequence is transmitted at a rate of typically 1–10 MHz, where higher rates imply a higher processing effort but promise more precise measurements. The PRN code is continuously repeated at intervals of a few milliseconds to seconds and facilitates measurements of the signal transmission time. In most GNSSs, the PRN sequence also serves as a unique fingerprint, which allows the receiver to distinguish individual satellites transmitting on the same frequency.

The Transit satellites used signals on two different frequencies (150/400 MHz) to cancel out ionospheric delays, a concept that was later inherited by GPS and the other GNSSs. Besides its main use as a navigation system, Transit also provided early contributions to geodesy and helped to establish a new global reference frame (Fig. 1.2).

Transit was decommissioned at the end of 1996 with the advent of GPS and its superior performance. And the equivalent Russian satellite Doppler systems have essentially been replaced by the Global'naya Navigatsionnaya Sputnikova Sistema (Russian Global Navigation Satellite System, [GLONASS](#)), the second fully operational GNSS. These new systems were based on the concept of range measurements rather than Doppler observations and used a different constellation design offering continuous coverage. These new concepts enabled a notable increase in accuracy as well as instantaneous positioning around the globe.

On top of the ranging code, the signal is also modulated with a low rate (e.g., 50 bits/s) navigation data stream (known as the broadcast navigation message) that provides information on the orbit of the transmitting satellite and the offset of its local clock from the GNSS system time.

The basic measurement made by a GNSS receiver is the time τ required for the GNSS signal to propagate from a satellite to the receiver. This can be obtained by tracking the PRN code modulation of the signal as illustrated in Fig. 1.3. Within the receiver, a local copy of the PRN sequence is generated, which is continuously compared and aligned with the signal received

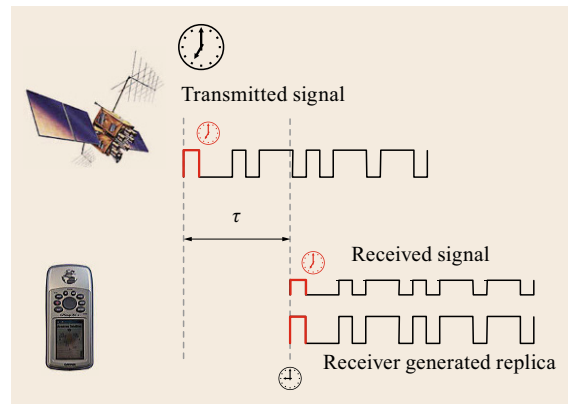


Fig. 1.3 Basic principle of pseudorange measurements

from the satellite. This *tracking loop* provides continuous measurements of the instantaneous code phase and hence the transmission time corresponding to the currently received signal (Chap. 13). By comparing this time with the local receiver time, the signal propagation time, and – upon multiplication by the speed of light – the distance or range from receiver to satellite are obtained.

Overall, the GNSS signals enable three basic types of measurements:

- **Pseudorange:** A measure of the difference between the receiver clock at signal reception and the satellite clock at signal transmission (scaled by the speed of light). Except for the asynchronicity of the two clocks and some other delays, the pseudorange measures the satellite–receiver distance, the precision of which is in the dm-range.
- **Carrier phase:** A measure of the instantaneous beat phase and the accumulated number of zero-crossings obtained after mixing with a reference signal of the nominal frequency. Changes in carrier phase over time reflect the change in (pseudo)range but are substantially (≈ 2 orders) more precise. In case of interrupted tracking the accumulated cycle count is lost and the carrier-phase measurements exhibit a cycle slip.
- **Doppler:** The change in the received frequency caused by the Doppler effect is a measure of the range-rate or line-of-sight velocity.

Pseudorange, carrier-phase, and Doppler observations provide the basic measurements for computing position and velocity as well as the offset of the receiver time with respect to the GNSS system time scale.

They are complemented by information on the orbit and clock offsets of the individual GNSS satellites, which is transmitted as part of the broadcast navigation message and allows the receiver to compute the position and velocity of the transmitting satellite at the signal transmission time. To provide such information with adequate accuracy, the GNSS operator must be able to determine and to predict the satellite orbit (Chap. 3) ahead of time, so that it can be uploaded to the satellite for subsequent broadcasting to the users. Likewise GNSS relies on highly stable onboard clocks, whose time offsets can be accurately predicted. Rubidium or cesium atomic frequency standards or even hydrogen masers are used for this purpose (Chap. 5), which deviate by only 10^{-13} to 10^{-14} from their nominal frequency over time scales of a day.

Before addressing pseudorange- and carrier-phase-based positioning in more detail, we first discuss the basic principles of range-based positioning using distance measurements.

1.2.2 Range-Based Positioning

As mentioned, the basic measurement made by a GNSS receiver is the time τ_r^s required for the GNSS signal to propagate from a satellite antenna s to a receiver antenna r . Since the signal travels at the speed of light, c , this time interval can be converted to a distance or range, simply by multiplying it by c

$$\rho_r^s(t) = c\tau_r^s. \quad (1.1)$$

Let us assume that the clock in the receiver is synchronized with the clock in the satellite, and that the atmosphere (ionosphere and troposphere), which slightly delays the arrival of the signal and about which we will talk later, does not exist. Furthermore, let us assume there is no measurement noise; that is, no random perturbation to the measurement, something that invariably affects all measurements to a greater or lesser degree. Under these ideal and simplified circumstances, the observation equation for the observed range takes the form

$$\begin{aligned} \rho_r^s(t) &= ||\mathbf{r}_r(t) - \mathbf{r}^s(t - \tau)|| \\ &= \left[(x_r(t) - x^s(t - \tau))^2 + (y_r(t) - y^s(t - \tau))^2 \right. \\ &\quad \left. + (z_r(t) - z^s(t - \tau))^2 \right]^{-\frac{1}{2}}, \end{aligned} \quad (1.2)$$

with $\mathbf{r}_r = (x_r, y_r, z_r)^\top$ being the unknown position vector of the receiver antenna (possibly moving and therefore a function of time) and $\mathbf{r}^s = (x^s, y^s, z^s)^\top$ that of the satellite (known from the navigation message transmitted by the satellite). Typically, both vectors are referred to an Earth-centered, Earth-fixed (ECEF) coordinate frame. Examples of such frames include versions of the World Geodetic System (WGS) 84 and the International Terrestrial Reference Frame (ITRF) (Chaps. 2 and 36).

With a single range measurement (1.2), we would know that the position of the receiver antenna must lie somewhere on a sphere, centered on the satellite, with a radius equal to the measured range; call it ρ_r^1 . If we simultaneously make a range measurement to a second satellite, then our receiver must also lie on a sphere, of radius ρ_r^2 , centered on this satellite. The two spheres will intersect, with the loci of intersection points forming a circle. Our receiver must lie somewhere on this circle, which is therefore called a line of position. A third simultaneous range measurement, ρ_r^3 , gives us a third sphere, which intersects the other two at just two points. One of these points can be immediately dismissed as being the location of our receiver, since it lies far out in space. So, the simultaneous measurement of the ranges to three satellites is sufficient to determine

a position fix in three dimensions – at least in principle (Fig. 1.4).

Computationally, the receiver position solution $(x_r, y_r, z_r)^\top$ of the simultaneous range equations

$$\begin{aligned}\rho_r^1 &= \sqrt{(x_r - x^1)^2 + (y_r - y^1)^2 + (z_r - z^1)^2} \\ \rho_r^2 &= \sqrt{(x_r - x^2)^2 + (y_r - y^2)^2 + (z_r - z^2)^2} \\ \rho_r^3 &= \sqrt{(x_r - x^3)^2 + (y_r - y^3)^2 + (z_r - z^3)^2}\end{aligned}\quad (1.3)$$

is usually obtained through an iterative linearization approach. Switching to vector formalism $\mathbf{p} = [\rho_r^1, \rho_r^2, \rho_r^3]^\top$ and dropping the indices, the system (1.3) of three non-linear range equations can be approximated to the first order as

$$\mathbf{p} = \mathbf{p}_0 + \mathbf{A} \Delta \mathbf{x}, \quad (1.4)$$

where \mathbf{p}_0 is the vector of computed range values based on given satellite coordinates (x^i, y^i, z^i) , $i = 1, 2, 3$, and an initial estimate $\mathbf{x}_0 = (x_{r,0}, y_{r,0}, z_{r,0})^\top$ of the receiver's position. Furthermore,

$$\mathbf{A} = \begin{pmatrix} \frac{\partial \rho_r^1}{\partial x_r} & \frac{\partial \rho_r^1}{\partial y_r} & \frac{\partial \rho_r^1}{\partial z_r} \\ \frac{\partial \rho_r^2}{\partial x_r} & \frac{\partial \rho_r^2}{\partial y_r} & \frac{\partial \rho_r^2}{\partial z_r} \\ \frac{\partial \rho_r^3}{\partial x_r} & \frac{\partial \rho_r^3}{\partial y_r} & \frac{\partial \rho_r^3}{\partial z_r} \end{pmatrix} \quad (1.5)$$

with $\partial \rho_r^i / \partial x_r = (x_{r,0} - x^i) / \rho_{r,0}^i$ ($i = 1, 2, 3$), is the design matrix, and $\Delta \mathbf{x} = \mathbf{x} - \mathbf{x}_0$ is the increment to the initial vector of receiver coordinates that is to be determined. Note that the matrix \mathbf{A} reflects the relative geometry of the satellites and the receiver. Solving for $\Delta \mathbf{x}$, we have

$$\Delta \mathbf{x} = \mathbf{A}^{-1}(\mathbf{p} - \mathbf{p}_0) = \mathbf{A}^{-1} \Delta \mathbf{p} \quad (1.6)$$

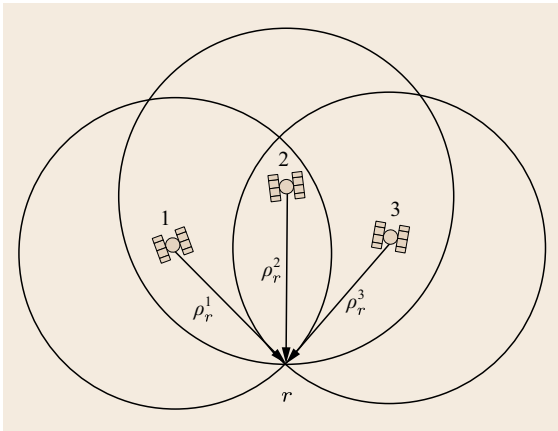


Fig. 1.4 Positioning through intersecting spheres

and then

$$\mathbf{x} = \mathbf{x}_0 + \Delta \mathbf{x}. \quad (1.7)$$

Depending on the closeness of the approximate ranges, \mathbf{p}_0 , to the measurements, \mathbf{p} , several iterations are, in general, required to arrive at final values for \mathbf{x} .

1.2.3 Pseudorange Positioning

So far we assumed that the clock in the GNSS receiver was synchronized with the clocks in the satellites. This assumption, however, is fallacious. When a GNSS receiver is switched on, its clock will in general be mis-synchronized with respect to the satellite clocks, by an unknown amount. Furthermore, the clocks in the satellites are synchronized with each other and to a master time scale, called the system time, only to within about a millisecond. The range measurements the receiver makes are biased by the receiver and satellite clock errors, dt_r and dt^s , and are therefore referred to as *pseudoranges*

$$p_r^s = \rho_r^s + c(dt_r - dt^s). \quad (1.8)$$

A timing error of a millisecond would result in an error in position of about 300 km, clearly an intolerable amount. It would be possible to better synchronize the satellite clocks by frequently sending them adjustment commands from the ground, but it has been found that clocks actually keep better time if they are left alone and the readings of the clock corrected. The GNSS operators monitor the satellite clocks and determine the offsets and drifts with respect to system time. These parameters are subsequently uploaded to the satellites and transmitted as part of a navigation message broadcast by the satellites. A GNSS receiver uses these satellite clock offset values to correct the measured pseudoranges.

However, we then still have the receiver clock error dt_r to deal with. Because of this error, the three spheres with radii equal to the measured pseudoranges corrected for the satellite clock offsets will not intersect at a common point. But, if the receiver clock error dt_r , can be determined, then the pseudoranges can be corrected and the position of the receiver determined. The situation, compressed into two dimensions, is illustrated in Fig. 1.5.

So now we actually have four unknown quantities or parameters in our pseudorange observation equation

$$p_r^s = \rho_r^s + c dt_r. \quad (1.9)$$

They are the three coordinates of the receiver antenna position (x_r, y_r, z_r) and the receiver clock offset dt_r .

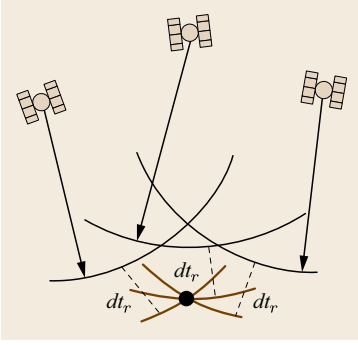


Fig. 1.5 Determination of receiver clock offset dt_r and true user position (intersection of *brown lines*) from the intersection of spheres centered on the satellites. Pseudoranges are shown by arcs of *black lines*

We thus need at least four simultaneous pseudoranges to estimate the three receiver coordinates and the receiver clock offset (measured in units of distance). With $\mathbf{x} = (x_r, y_r, z_r, dt_r)^T$ and the four-by-four design matrix

$$\mathbf{A} = \begin{pmatrix} \frac{\partial \rho_r^1}{\partial x_r} & \frac{\partial \rho_r^1}{\partial y_r} & \frac{\partial \rho_r^1}{\partial z_r} & 1 \\ \frac{\partial \rho_r^2}{\partial x_r} & \frac{\partial \rho_r^2}{\partial y_r} & \frac{\partial \rho_r^2}{\partial z_r} & 1 \\ \frac{\partial \rho_r^3}{\partial x_r} & \frac{\partial \rho_r^3}{\partial y_r} & \frac{\partial \rho_r^3}{\partial z_r} & 1 \\ \frac{\partial \rho_r^4}{\partial x_r} & \frac{\partial \rho_r^4}{\partial y_r} & \frac{\partial \rho_r^4}{\partial z_r} & 1 \end{pmatrix} \quad (1.10)$$

the same iterative procedure as described before can then be applied.

What if signals from more than four satellites are available? Because of the unmodeled errors (e.g., atmospheric delays) as well as the residual errors in the modeled terms, it is beneficial to use simultaneous pseudoranges to all m available satellites for estimating the receiver coordinates and clock offset. This requires use of a nonlinear least-squares (or related Kalman filter) estimation procedure (Chap. 22)

$$\Delta \mathbf{x} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \Delta \mathbf{p}, \quad (1.11)$$

where \mathbf{A} now has dimensions of $m \times 4$ and \mathbf{W} is a weight matrix, which reflects the uncertainty in the observations and any correlations that may exist among them. This weight matrix may be written as

$$\mathbf{W} = \mathbf{Q}_{pp}^{-1}, \quad (1.12)$$

in which \mathbf{Q}_{pp} is the covariance matrix of the pseudorange errors. In general, the solution of a nonlinear problem must be iterated to obtain the result. However, if the linearization point is sufficiently close to the true solution, then only one iteration is required (Chaps. 21 and 22). The processing of measurements

can take place in real time within the receiver or the raw measurements can be stored for postprocessing by a standalone computer.

1.2.4 Precision of Position Solutions

The precision with which the receiver's coordinates and clock offset can be determined is described by the covariance matrix of the solution $\Delta \mathbf{x}$. This covariance matrix, denoted as \mathbf{Q}_{xx} , follows, with (1.12), from applying the error propagation law, also known as the variance propagation law, to (1.11) as

$$\begin{aligned} \mathbf{Q}_{xx} &= [(\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W}] \\ &\quad \cdot \mathbf{Q}_{pp} \cdot [(\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W}]^T \\ &= (\mathbf{A}^T \mathbf{Q}_{pp}^{-1} \mathbf{A})^{-1}. \end{aligned} \quad (1.13)$$

The diagonal elements of this matrix are the estimated receiver coordinate and clock-offset variances, and the off-diagonal elements (the covariances) indicate the degree to which these estimates are correlated. This equation represents a fundamental relationship widely used for actual measurement analysis as well as for experiment and system design studies. It allows one to examine the effect a particular design (through design matrix \mathbf{A}) or measurement capability (through measurement covariance matrix \mathbf{Q}_{pp}) will have on specified parameters without actually making any measurements.

In GNSS-related studies, for example, we might use the equation to answer a variety of questions:

- What is the behavior of the estimated parameter covariance matrix as a function of particular satellite configurations?
- How do various model errors propagate into the receiver coordinates as a function of satellite configurations?
- What is the tolerance value that a particular model error should not exceed to achieve a desired positioning accuracy?

User Equivalent Range Error

Such analyses simplify if we assume that the measurement and residual model errors are uncorrelated and the same for all observations with a particular standard deviation (σ). Then $\mathbf{Q}_{pp} = \sigma^2 \mathbf{I}_m$ (in which \mathbf{I}_m is the identity matrix of order m) and the expression for the covariance matrix of \mathbf{x} simplifies to

$$\mathbf{Q}_{xx} = \sigma^2 (\mathbf{A}^T \mathbf{A})^{-1}. \quad (1.14)$$

When we combine satellite clock and ephemeris error, atmospheric error, receiver noise, and multipath – all expressed in units of distance – we obtain a quantity

known as the total *user equivalent range error* (**UERE**), which we can use for σ . UERE can further be divided into two parts [1.6], namely:

- The *signal-in-space (user) range error* (SISRE or SISURE), which comprises errors related to the space and control segment (primarily broadcast satellite orbit and clock errors), and
- The *user equipment error* (UEE), which includes the remaining contributions specific to the user's receiver and environment.

The total UERE can then be written as

$$\text{UERE} = \sqrt{\text{SISRE}^2 + \text{UEE}^2}. \quad (1.15)$$

An overview of key contributions to the UERE budgets is provided in Table 1.1. The given values are mainly for illustration, since it is difficult to provide universally valid error bounds in most cases. Among others, UEE contributions may differ widely among individual receivers and sites. In the case of GPS single-frequency positioning, the total UERE is typically in the neighborhood of a few meters, with the actual value dominated by ionospheric and multipath effects [1.7]. Dual-frequency positioning, with the capability to remove almost all of the ionospheric delay from the pseudorange observations, can result in even smaller UEREs.

Dilution of Precision

A simple scalar indicator of the overall quality of the least-squares solution is given by the square root of the sum of the parameter estimate variances

$$\begin{aligned} \sigma_G &= \sqrt{\sigma_E^2 + \sigma_N^2 + \sigma_U^2 + \sigma_{dt}^2} \\ &= \sigma \text{tr}\{(\mathbf{A}^\top \mathbf{A})^{-1}\}, \end{aligned} \quad (1.16)$$

Table 1.1 Representative magnitudes of individual contributions to the GNSS user equivalent range error (after [1.8–10]) for estimates of the individual contributions

Error source	Contribution 1 σ (m)
SISRE	
Broadcast satellite orbit	0.2–1.0
Broadcast satellite clock	0.3–1.9
Broadcast group delays	0.0–0.2
UEE	
Unmodeled ionospheric delay	0–5
Unmodeled tropospheric delay	0.2
Multipath	0.2–1
Receiver noise	0.1–1
UERE	0.5–6

in which σ_E^2 , σ_N^2 , and σ_U^2 are the variances of the east, north, and up components of the receiver position estimate, respectively, and σ_{dt}^2 is the variance of the estimated receiver clock offset. If the solution algorithm is parameterized in terms of geocentric Cartesian coordinates, it is a straightforward procedure to transform the solution variance matrix to the local coordinate frame to get the north, east, and up components.

The elements of matrix $\mathbf{A}^\top \mathbf{A}$ are a function of the receiver–satellite geometry and as the trace of its inverse is typically greater than 1, it amplifies σ , or dilutes the precision, of the position determination. This scaling factor is therefore usually called the geometric dilution of precision (GDOP). The GDOP becomes the position-DOP (**PDOP**), if one leaves out the contribution of the receiver clock, and it further reduces to the horizontal-DOP (**HDOP**), if one also leaves out the contribution of the up-component. In a likewise manner one can obtain the vertical-DOP (**VDOP**).

It turns out that DOP values depend on the volume of the polyhedron formed by the tips of the receiver–satellite unit vectors. The larger the volume, the smaller the DOPs. If the tips lie in a plane, the DOP factors are infinitely large. In fact, no position solution is possible with this receiver–satellite geometry as the matrix $\mathbf{A}^\top \mathbf{A}$ is singular: the solution cannot distinguish between an error in the receiver clock and an error in the position of the receiver. DOP values are smaller and hence solution errors are smaller when the satellites used in computing the solution are spread out in the sky.

High DOP values can sometimes occur even for all-in-view receivers operating at mid-latitudes. In some environments, such as heavily forested areas or urban canyons, a GNSS receiver's antenna may not have a clear view of the whole sky because of obstructions. If it can only receive GNSS signals from a small region of the sky, the DOPs will be large, and position accuracy will suffer. Being able to track more satellites can help in such situations, and a multi-GNSS receiver may provide acceptable accuracies. New receiver technology permitting use of weaker GPS signals, even those present inside buildings, will also be beneficial.

While DOP and UERE are highly useful concepts to understand GNSS positioning errors and their dependence on the geometric distribution of tracked satellites as well as individual pseudorange error sources, readers should keep in mind that the common rule-of-thumb

$$\text{Navigation error} = \text{DOP} \times \text{UERE} \quad (1.17)$$

is only a rough approximation and limited to random error propagation. Efforts to better account for system-

atic and random errors and to arrive at a more realistic description of real positioning errors are, for example, presented in [1.11].

1.2.5 GNSS Observation Equations

So far, we have been working with a rather simplified version of the pseudorange observation equation (1.8) and (1.9). In actuality, however, there are a number of additional error sources that must be taken into consideration and modeled in the observation equation. We therefore have to modify the pseudorange observation equation as follows

$$p_r^s = \rho_r^s + c(dt_r - dt^s) + T_r^s + I_r^s + \epsilon_r^s. \quad (1.18)$$

Here, dt_r and dt^s are the receiver and satellite clock offset from GNSS system time as before, T_r^s is the neutral atmosphere (troposphere) propagation delay, I_r^s is the ionospheric propagation delay, and ϵ_r^s represents unmodeled errors including receiver noise, multipath, and other small effects (Chaps. 13–15). For a more detailed discussion of the basic pseudorange observation equation, see Chap. 19.

The carrier-phase observation equation is similar to the pseudorange equation

$$\varphi_r^s = \rho_r^s + c(dt_r - dt^s) + T_r^s - I_r^s + \lambda M_r^s + \epsilon_r^s. \quad (1.19)$$

In addition to the previously defined terms, λ is the carrier wavelength, $M_r^s = N_r^s + \delta_r - \delta^s$ is the sum of the integer carrier-phase ambiguity N_r^s (in cycles) and the instrumental receiver and satellite phase delays $\delta_r - \delta^s$ (in cycles), ϵ_r^s represents unmodeled phase errors including receiver noise, multipath, and other small effects.

Several terms in the carrier-phase observation equation are nominally identical to those in the pseudorange observation equation, including the geometric range, receiver and satellite clock offsets, and the tropospheric propagation delay. The magnitude of the ionospheric term in both equations is the same, however its sign is negative in the carrier-phase equation. This relates to the fact that the phase of the carrier is advanced during the signal's passage through the ionosphere, as opposed to the pseudorange, which suffers a delay. And the ionospheric phase advance is frequency dependent like the pseudorange delay. For a more detailed discussion of the basic carrier-phase observation equation, see Chap. 19.

The receiver or postprocessing software will use the above forms of the observation equations to compute the receiver coordinates or any other relevant GNSS parameters. Models exist for describing the neutral atmosphere propagation delay, and the ionospheric propagation delay can be corrected using the coefficients of a model included in the broadcast navigation message or by combining simultaneous pseudorange or carrier-phase measurements made on two transmitted frequencies (Chaps. 38 and 39).

1.3 Modeling the Observations

To accurately determine the receiver coordinates or any other relevant GNSS parameters, we must model the right-hand sides of the observation equations (1.18) and (1.19) to match as accurately as possible the receiver's observations of the satellites. This requires knowledge of the satellites' positions at the time of signal transmission, the offset of their clocks, the atmospheric propagation delays, ambiguities for the carrier-phase observations, plus perhaps some smaller contributions such as instrumental delays. If the available information is not sufficiently accurate, we may be able to estimate residual effects from the observations themselves and so better match the right-hand and left-hand sides of the equations.

In the next few sections, we will overview some of the information required for a receiver or external software to process GNSS observations. More detailed descriptions are provided in subsequent chapters of the Handbook.

1.3.1 Satellite Orbit and Clock Information

The predicted position of a satellite's antenna phase center can be derived from the Keplerian-like orbit parameters given in the broadcast navigation message. The predicted offset of the active satellite clock from system time is also derived from the navigation message. The combined accuracy of these terms, SISRE, is typically 0.5–2 m for current global and regional navigation satellite systems. For GPS, the Global Positioning Systems Directorate has reported that the annual root-mean-square signal-in-space range error (SISRE) across all healthy satellites dropped from 1.6 m in 2001 to 0.7 m in 2014 [1.12] and further improvement has been made after replacing the old Block IIA satellites with the latest generation of Block IIF satellites in early 2016. SISRE values of about 0.5 m can also be expected for Galileo based on the performance of the preoperational constellation in the 2015/2016 timeframe. The

performance of GLONASS, the second fully operational GNSS, is presently about three times worse but likewise expected to improve with the ongoing modernization of the space and ground segments.

Significantly more accurate satellite orbit and clock information is available from various sources, such as satellite-based augmentations systems, the international GNSS service, and private service providers, for both real-time and postprocessing applications (Chaps. 3, 33, and 34).

1.3.2 Atmospheric Propagation Delay

GNSS signals travel through the Earth's atmosphere to receivers on or near the ground (Chap. 6). The signals are refracted, changing their velocity – both speed and direction of travel. Measured pseudoranges and carrier phases are biased by meters to tens of meters. The biases are determined by the integrated effect of the refractive index – the ratio of the speed of propagation of an electromagnetic wave in a vacuum to that in a medium – all along the signal raypath. In general, the refractive index will be a function of the characteristics of the medium including the type and densities of the medium's constituents as well as the frequency of the wave and possibly external factors such as ambient magnetic fields.

The atmospheric propagation delays can be separated into those due to the electrically neutral atmosphere and those due to the ionosphere (Fig. 1.6).

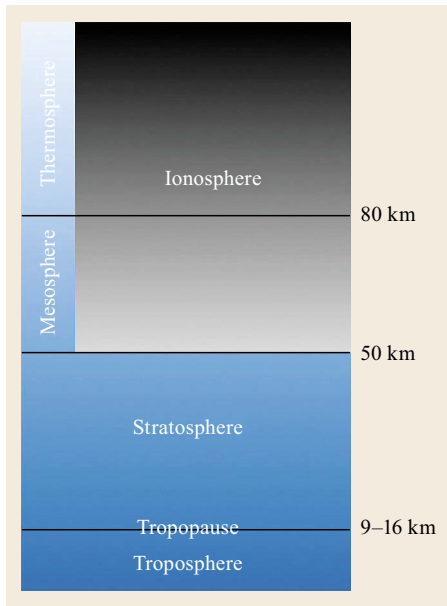


Fig. 1.6 Structure of the Earth's atmosphere

Neutral Atmosphere

The neutral atmosphere is that part of the atmosphere that is electrically neutral and stretches from ground level up to a height of 50 km and beyond (Chap. 6). It is what we colloquially refer to as the air. Air is made up of nitrogen, oxygen, carbon dioxide, as well as some other atoms and molecules including water vapor. With the inclusion of water vapor, we refer to the medium as moist air. The refractive index n of a parcel of moist air is a function of its temperature, the partial pressures P_d of the dry constituents (nitrogen, oxygen, etc.) and the partial pressure e of water vapor

$$n = n(T, P_d, e) . \quad (1.20)$$

Note that air is essentially a nondispersive medium, with n independent of frequency throughout most of the radio spectrum and including the frequencies used by all GNSSs. This also means that the effect on pseudoranges is identical to that on carrier-phase measurements.

At sea level, values of the refractive index of air are close to 1.0003, becoming smaller with increasing height. A more useful quantity is refractivity, $N = 10^6(n - 1)$, with sea level values near 300. Since the bulk of the neutral atmospheric effect occurs in the lowest-most part of the atmosphere – the troposphere – the effect is often termed the tropospheric propagation delay.

How much delay is imparted to a GNSS signal by the neutral atmosphere? This depends on the location and height of the receiver as well as weather conditions and also the elevation angle (and, to a lesser degree, azimuth) at which the signal arrives at the receiver. The total delay experienced by the received signal is referred to as the slant delay. It is modeled by mapping the delay of a hypothetical signal arriving from directly overhead – the zenith direction – to the actual signal slant path using a mapping function (also called an obliquity factor). Typically, at sea level, zenith delay is about 2.4 m, rising to more than 24 m at an elevation angle of 5° .

Various neutral atmosphere delay models exist. Many receivers use the Radio Technical Commission for Aeronautics (RTCA) MOPS (Minimum Operational Performance Standards model) [1.13] (a slightly simplified version of the University of New Brunswick's UNB3 model, the predecessor to UNB3m [1.14, 15]) mandated for use by satellite-based augmentation system (SBAS) user equipment. Several more sophisticated models exist (Chap. 6). No model is perfect, though, and there is some advantage in estimating residual delays from the GNSS data itself.

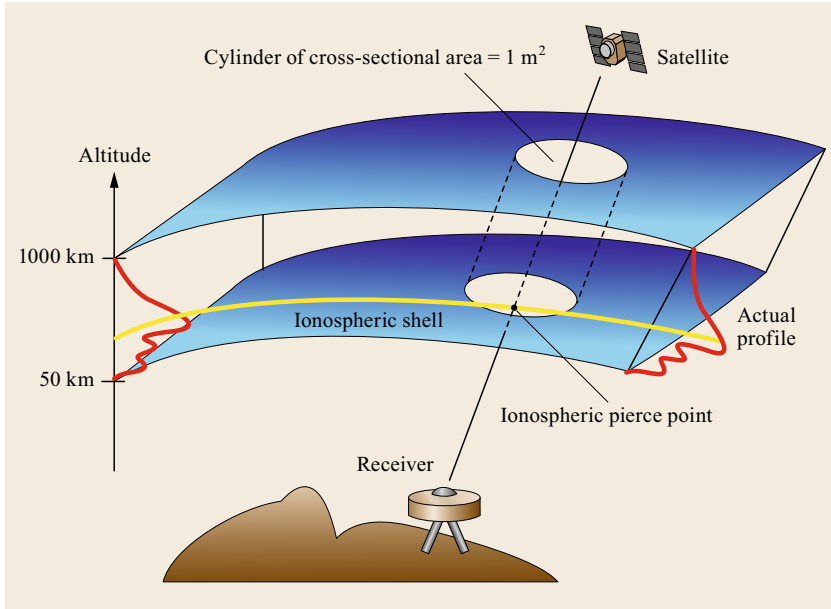


Fig. 1.7 The thin-shell approximation of the ionosphere (after [1.16])

Ionosphere

The ionosphere is that region of the Earth's atmosphere in which ionizing radiation (principally from solar extreme ultraviolet (EUV) and x-ray emissions) cause electrons to exist in sufficient quantities to affect the propagation of radio waves (Sect. 6.3). It extends from about 50–1000 km or more, above which we have the plasmasphere (also known as the protonosphere).

The ionosphere is a dispersive medium for radio waves: the refractive index is a function of frequency. It also means that pseudorange and carrier-phase observations are affected differently, with a phase refractive index and a pseudorange (group delay) refractive index. This results in an ionospheric delay for pseudoranges (increased value) and an ionospheric phase advance for carrier-phase measurements (decreased value). In addition to the radio frequency of the signal, the refractive indices are a function of electron density. The Earth's magnetic field also plays a minor role.

Again, it is the integrated effect of ionospheric refractive index all along the raypath that determines the pseudorange delay and the carrier-phase advance. It turns out that, to first order, the magnitudes of the pseudorange delay and the carrier-phase advance are the same; they just differ in sign. To an excellent approximation, the magnitude, I , in m is given by

$$I = 40.3 \frac{\text{TEC}}{f^2}, \quad (1.21)$$

where **TEC** is the total electron content and is the total number of electrons in a column of one-meter-square

cross-section centered on the signal raypath and stretching from the receiver to the satellite. In this equation, TEC is given in electrons/m², and the frequency, f , is given in Hz. Typical TEC values measured near the Earth's surface range from about 10^{16} to 10^{19} with the actual value depending on geographic location, local time, season, solar EUV flux, and magnetic activity. And, as with neutral atmosphere delays, we have slant and zenith delays, corresponding to slant and zenith TEC.

Since the ionospheric effect is to a very good approximation inversely proportional to the square of the frequency, by linearly combining simultaneous measurements (either pseudoranges or carrier phases) on two frequencies such as the GPS L1 and L2 frequencies, an observable virtually free of ionospheric effects can be constructed and used for position determinations (Sect. 20.2.3). This approach does require, however, a dual- or multifrequency receiver.

If only single-frequency observations are available, then the sum of the pseudorange and phase measurement can be taken to eliminate the ionospheric delay, or alternatively a model can be used to account for the ionospheric biases as much as possible. The GPS navigation message includes values of the parameters of a simple ionospheric model known as the broadcast or *Klobuchar* model [1.17], named after its developer *Jack Klobuchar*. This model permits an estimate of the zenith ionospheric delay to be computed at a receiver's location at a particular time of day and is driven by recent solar conditions as interpreted by the GPS control

segment. The zenith delay is then mapped into a slant delay assuming a thin shell model for the ionosphere, where all of the electron content is assumed to occur in a shell at a particular height above the Earth's surface (Fig. 1.7). The broadcast model assumes a shell height of 350 km.

A Klobuchar-like model is also used within the BeiDou system, while a version of the NeQuick model [1.18] has been selected as an alternative for the Galileo system. NeQuick describes the 3-dimensional

electron density distribution as a function of a small set of ionospheric activity parameters provided in the navigation message and an extensive set of static, seasonal coefficients. The ionospheric slant delay is then obtained by integrating the electron density along the ray path between satellite and receiver. While computationally more demanding, the NeQuick model requires a smaller set of broadcast parameters and achieves an improved overall correction performance compared to the GPS Klobuchar model [1.10].

1.4 Positioning Modes

There are a variety of GNSS positioning (and navigation) modes of differing degrees of complexity and precision and accuracy. These range from the standard single-frequency pseudorange-based approach used by most consumer receivers including those in mobile phones to high-integrity methods for safety-of-life applications to sophisticated multifrequency carrier-phase-based techniques capable of centimeter to subcentimeter accuracies for demanding applications like machine control and scientific studies. In this section, we will briefly outline some of the approaches with deeper discussions appearing in later chapters (Chaps. 21, 23, 26, and 35).

1.4.1 Precise Point Positioning

Precise point positioning (PPP) is an advanced version of the single-point positioning (SPP) technique that we discussed earlier. PPP (Chap. 25) uses carrier-phase measurements as the primary observable with pseudorange measurements playing a secondary role. The PPP processing algorithm is similar to that for pseudorange-measurement positioning, except that effects down to the centimeter-level and lower must be modeled or estimated. This means that satellite constellation precise orbits and clock offsets (Chap. 34) must be used such as those provided by the International GNSS Service (IGS). Typically, a dual-frequency GNSS receiver is used with dual-frequency code and phase measurements linearly combined to remove first-order ionospheric effects. The carrier-phase ambiguities are estimated (resolving them to integer values if possible) as well as residual tropospheric propagation delay after applying an a priori model. Subtle effects, such as Earth tides, ocean tide loading, satellite and receiver antenna offsets, and carrier-phase windup are also modeled.

The performance of PPP can be measured in terms of accuracy, precision, convergence period (the time

required for a position solution to converge below a certain accuracy threshold), availability, and integrity. Best effort PPP is generally bias free, so there is little difference between accuracy and precision statistics. PPP can provide few-centimeter-level 1-sigma accuracies in each coordinate (north, east, and up) for a static site after convergence whereas decimeter-level accuracies can be achieved for a moving platform or a static site with data processed in *kinematic mode* (independent epoch-by-epoch position fixes). Accuracies, although good, may not be sufficient for some applications. The convergence period for achieving a decimeter-level solution is typically up to about 30 min under normal conditions with some newer procedures (multifrequency, multi-constellation) achieving convergence periods of 20 min or less. Continuous availability of accurate PPP fixes depends on the environment. Signal blockages by trees and buildings can reduce availability but multiconstellation observations can be a big help in this regard. Integrity measures for PPP are currently limited.

PPP can be used for processing data from either static (stationary) sites or kinematic (moving) platforms

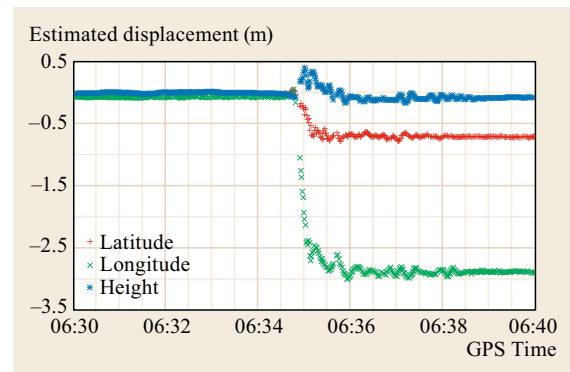


Fig. 1.8 Estimate of co-seismic displacement at IGS station CONZ following the 8.8 magnitude Chilean earthquake of 27 February 2010 (after [1.19])

and its uses include establishing and updating reference-station coordinates for crustal-deformation monitoring (Fig. 1.8), precise orbit determination of low-Earth-orbiting satellites, ocean buoy positioning for tsunami detection with main commercial applications in precision farming, seafloor mapping, marine construction, and airborne mapping. Application of PPP is expanding to atmosphere remote sensing, precise time transfer, land surveying, construction, and military uses.

1.4.2 Code Differential Positioning

The advantage of differential positioning (Chap. 26) over SPP is that with differential techniques certain effects are eliminated or largely reduced (e.g., orbit errors and atmospheric delays with dependence on the spatial correlation). There are two basic kinds of code differential positioning: measurement domain (typically covering a local/regional area and known as differential GPS/GNSS or DGPS/DGNSS) and state-space domain (typically covering a wide area and known as wide-area GPS/GNSS or WADGPS/WADGNSS). The best example of a measurement-domain DGPS is that implemented by coast guard agencies around the world for maritime navigation (Sect. 29.4). The best example of state-space-domain DGPS is that of SBASs, the first of which was the US Federal Aviation Administration's (FAA's) Wide Area Augmentation System (WAAS). SBAS is discussed in Sect. 1.5 and in detail in Chap. 12.

The measurement domain techniques provide composite measurement corrections to the user without estimating individual error components whereas state-space domain techniques provide individual error corrections such as satellite orbit and clock and ionospheric propagation delay. These corrections are determined at reference stations and transmitted to users using radio beacons.

The combined corrections account for navigation message satellite orbit and clock error, tropospheric propagation delay, ionospheric propagation delay and, in the past, GPS Selective Availability. A user requires a GNSS receiver with an integrated beacon receiver or a separate beacon receiver connected to the GNSS receiver by a serial communications link such as RS-232. The corrections are datum dependent. In North America, user-computed positions will be in the North American Datum (NAD) 83 system, not WGS 84.

Position accuracy generally degrades with increasing distance from the beacon transmitter site. Official accuracy is stated as 10 m (horizontal at 95%) within the coverage area but typically, the error of a DGPS position is 1–3 m. Accuracy will be affected by user

multipath and DOP. The error is often seen as a bias in positioning, resulting in a position offset. The scatter of the coordinates is likely to remain close to constant. A general rule of thumb is an additional 1 m error per 100 km. However, accuracy is worse during strong ionospheric disturbances due to gradients.

1.4.3 Differential Carrier Phase

Differential carrier-phase positioning is a classic technique dating to the early 1980s. The procedure combines data from one (or more) reference stations with user data (Chap. 26). Observations on the same satellite at the same epoch are differenced between receivers (single difference) and then single differences are differenced between pairs of satellites (double difference). This procedure eliminates residual satellite and receiver clock errors, and reduces satellite orbit error and atmospheric propagation delay errors (Chap. 20). Accuracies at the decimeter level and better can be obtained. In principle, a similar approach could be taken using pseudoranges (alone) but with much lower resulting accuracies. See Chap. 26 for a more detailed description.

Single Differencing

Let us take a look at the differencing operations in detail starting with single differencing. Differencing carrier-phase measurements on one satellite s from two receivers, 1 and 2, gives us

$$\varphi_{12}^s = \rho_{12}^s + cdt_{12} + T_{12}^s - I_{12}^s + \lambda M_{12}^s + \epsilon_{12}^s, \quad (1.22)$$

where $\varphi_{12}^s = \varphi_2^s - \varphi_1^s$ is the between-receiver single-difference carrier phase, with a similar notational convention for the terms on the right-hand side (1.22). Note that the satellite clock error has disappeared because it is identical for both receivers. Likewise, the satellite phase bias disappeared from $M_{12}^s = \delta_{12} + N_{12}^s$.

Differencing measurements made on two satellites, s and t , with receiver 1 gives us

$$\varphi_1^{st} = \rho_1^{st} + cdt_1^{st} + T_1^{st} - I_1^{st} + \lambda M_1^{st} + \epsilon_1^{st}, \quad (1.23)$$

where $\varphi_1^{st} = \varphi_1^t - \varphi_1^s$ is the between-satellite single-difference carrier phase. Note that the receiver clock error has disappeared because it is identical for measurements made on all satellites at the same epoch. Likewise, the receiver phase bias disappeared from $M_1^{st} = \delta^{st} + N_1^{st}$.

Double Differencing

If we difference carrier-phase measurements between receivers and then difference the resulting values between satellites, we have the double-difference observ-

able

$$\varphi_{12}^{st} = \rho_{12}^{st} + T_{12}^{st} - I_{12}^{st} + \lambda N_{12}^{st} + \epsilon_{12}^{st}, \quad (1.24)$$

where $\varphi_{12}^{st} = \varphi_2^{st} - \varphi_1^{st} = \varphi_{12}^t - \varphi_{12}^s$. Note that both receiver and satellite clock error have disappeared as well as the corresponding instrumental phase biases. As a result, the real valued parameters, M_{12}^s in (1.22) and M_1^{st} in (1.23), have been replaced by the integer parameter N_{12}^{st} in (1.24).

In processing double differences (between a pair of receivers and a pair of satellites), we must model the double-difference geometric range, the double-difference tropospheric delay, the double-difference ionospheric effect (which can be neglected in case of sufficiently short baselines), and the double-difference phase ambiguity, here now an integer. Least squares or a Kalman filter approach is used to estimate the coordinates of the user receiver along with the *nuisance* parameters in a similar approach to that used for processing undifferenced pseudorange or carrier-phase observations. Redundancy in the system of observation equations is used to check the validity of the assumed models (Chaps. 22 and 24).

The integer ambiguities N_{12}^{st} for all satellite pairs in an observation set are estimated together with the coordinates of the user receiver and (optionally) a residual tropospheric delay error. The ambiguities are initially estimated as floating-point numbers (the so-called *float solution*); procedures are available for fixing some or all of the ambiguities at their correct integer values (the so-called *fixed solution*) (Chap. 23). Successful ambiguity resolution depends on several factors including baseline length (shorter is better), the number of satellites in view (more is better), continuous tracking of satellites, low dilution of precision values, the degree of multipath (less is better), the number of frequencies observed (two are better than one), and length of observing session (longer is better). With modern receivers and techniques, ambiguity resolution can be carried out with just a few tens of seconds of observations, even if the user receiver is in motion (*on the fly*). Fixed solutions generally provide more accurate results. A common ambiguity fixing technique is LAMBDA (Least-squares Ambiguity Decorrelation Adjustment) [1.20] (Chap. 23).

Real-Time Kinematic Positioning

In real-time kinematic (RTK) positioning, a GNSS reference station transmits carrier-phase and pseudorange data over a radio link to a roving station. Either single- or dual-frequency GNSS receivers can be used, with the dual-frequency systems typically affording faster ambiguity resolution and higher positioning accuracies over

longer distances. The receivers must incorporate data radios (or be wired to external radios), typically operating in the very high frequency (VHF, 30–300 MHz) or ultrahigh frequency (UHF, 300 MHz and 3 GHz) parts of the radio spectrum. The reference station transmits pseudorange and carrier-phase measurements and ancillary data. Radio Technical Commission for Maritime Services (RTCM) SC-104 2.x or 3.x [1.21] data protocols are typically used although proprietary data formats also exist.

Transmission modes vary and include narrow-band frequency modulation (FM) with frequency-shift-keying and packetized data transmission. Both 2 and 35 W transmitters are commonly available and for these licensing is typically required. However, license-free, low-power transmitters can also be used. In any case, VHF/UHF data links are limited to line of sight and transmitting and receiving antennas should be as high as possible. The maximum theoretical range is given by [1.22, 23]

$$d(\text{km}) = 3.57\sqrt{k} \left[\sqrt{h_t(\text{m})} + \sqrt{h_r(\text{m})} \right]. \quad (1.25)$$

with h_t and h_r the height of transmitter and receiver, respectively, and where k varies with refractivity (typically between 1.2 and 1.6); for example, for a transmitting antenna at 30 m above the terrain and a receiving antenna at 2 m, the maximum range is 28 km. Any obstructions along the propagation path will affect the signal's range. Whether or not a received signal can be successfully used depends on several factors including receiver sensitivity.

Reference station data can also be transmitted via the Internet using, e.g., the Network Transportation of RTCM Internet Protocol (NTRIP; [1.24]) and accessed by a hardwire or wireless link such as a mobile phone.

To reduce latency effects, transmission protocols typically use data differences with reconstruction of reference station data at the rover. Cycle slips must be detected and corrected in real time and ambiguities must be resolved quickly, even if the receiver is in motion (*on the fly*). Several techniques exist. Use of GLONASS data in addition to GPS data can result in greater availability, faster ambiguity fixing, and higher positioning accuracies.

PPP-RTK, a combination or merging of PPP with state-space RTK can have significant advantages in ambiguity resolution, in convergence time, and in accuracy and practical systems have been implemented by industry and research organizations (Chaps. 25 and 26). All individual GNSS error components, derived from the RTK monitoring network, are determined and delivered using state-space representation (SSR). These include orbits, clocks, code (pseudorange) biases, ionosphere

(for single-frequency receivers), troposphere, and carrier-phase biases. In principle, the concept can be applied to small, regional, and global networks. RTCM

and IGS have active committees developing standards for PPP-RTK and real-time PPP for delivering prototype research and commercial services.

1.5 Current and Developing GNSSs

GNSSs and regional navigation satellite systems (RNSSs) commonly consist of three components:

- The *space segment* comprises a constellation of satellites orbiting above the Earth's surface that transmit ranging signals on at least two frequencies in the microwave part of the radio spectrum.
- The *control segment* is responsible for maintaining the health of the system by monitoring the broadcast signals and computing and uploading to the satellites required navigation data. It consists of a group of globally (or locally)-dispersed monitoring stations, ground antennas for communicating with the satellites, and a master control station with a backup facility at a different location.
- The *user segment* consists of GNSS receiving equipment both civil and military. This includes receivers on the ground, at sea, in the air, and even in space.

GNSS constellations typically adopt a specific orbital configuration: **MEO** (medium-altitude Earth orbit) satellites for global coverage; **IGSOs** (inclined geosynchronous orbits) and **GEOs** (geostationary orbits) as supplements in regional systems.

MEO satellites are often evenly distributed in inclined near-circular orbits overequally spaced orbital planes, forming a constellation known as a *Walker* constellation [1.25]. The geometry of a specific Walker constellation is described by the triplet $t/p/f$, where t denotes the total number of satellites, p the number of equally spaced planes, and f the phase difference between the adjacent orbital planes. To determine the angle between satellites in adjacent planes, the parameter f should be multiplied by $360^\circ/t$. A simple Walker constellation of eight satellites in two orbital planes is illustrated in Fig. 1.9.

Currently, there are six GNSSs/RNSSs in operation. The four GNSSs are: GPS (US), GLONASS (Russia), BeiDou (China), and Galileo (EU); and the two RNSSs: QZSS (Japan) and IRNSS/NavIC (India). For an overview summary, see Table 1.2.

1.5.1 Global Navigation Satellite Systems

GPS

The Global Positioning System (GPS; Chap. 7) is the US GNSS, which provides free positioning and timing services worldwide. It was originally developed for the US military and was made free for civilian purposes very early in the experimental phase of GPS. The launch of the first Block I Navstar GPS satellite occurred on 22 February 1978, followed by the declaration of the Initial Operating Capability in December 1993 with 24 operational satellites in orbit, and the Full Operational Capability in June 1995. GPS is maintained by the US government and is freely accessible by anyone with a GPS receiver. GPS provides two different positioning services: the Precise Positioning Service (PPS) on the GPS L1 (1575.42 MHz) and L2 (1227.6 MHz) frequencies both containing an encrypted precision (P) code ranging signal (known as the Y-code) with a navigation data message for authorized users, and the Standard Positioning Service (SPS) on the GPS L1 frequency containing a coarse/acquisition (C/A) code and a navigation data message for civilian users. The GPS modernization program began in 2005 with the launch of the first IIR-M satellite. Since that moment on, two new signals have been transmitted: L2C for civilian users and a new military signal (M-code) at the L1 and L2 frequencies to provide better jamming resistance than the Y-code. Moreover, a new radio frequency link L5 (1176.45 MHz) for civilian users has been introduced. This signal, available since

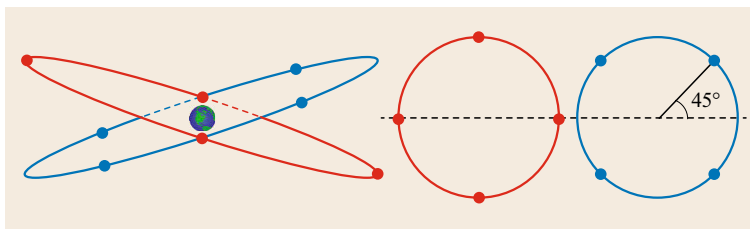







Fig. 1.9 Schematic illustration of an 8/2/1 Walker constellation

Table 1.2 An overview of the global and regional satellite-based navigation systems. Logos reproduced with permission of IAC PNT and FGUP TSNIIMASH (GLONASS), China Satellite Navigation Office (BeiDou), QZS System Services Inc. (QZSS), and Indian Space Research Organization (IRNSS/NavIC). The Navstar GPS logo is in the public domain

System	GPS	GLONASS	BeiDou	Galileo	QZSS	IRNSS/NavIC
						
Orbit	MEO	MEO	MEO, IGSO, GEO	MEO	IGSO, GEO	IGSO, GEO
Nominal number of satellites	24	24	27, 3, 5	30	3, 1	4, 3
Constellation	6 planes 56° inclination	Walker (24/3/1) 64.8° inclination	Walker (24/3/1) 55° inclination	Walker (24/3/1) 56° inclination	IGSOs with 43° inclination	IGSOs with 29° inclination
Services	SPS, PPS	SPS, PPS	OS, AS, WADS, SMS	OS, CS, PRS	GCS, GAS, PRS, EWS, MCS	SPS, RS
Initial service	Dec 1993	Sep 1993	Dec 2012	2016/2017 (planned)	2018 (planned)	2016 (planned)
Origin	USA	Russia	China	Europe	Japan	India
Coverage	Global	Global	Global	Global	East Asia Oceania region	$-30^\circ < \phi < 50^\circ$ $30^\circ < \lambda < 130^\circ$
Frequency (MHz)	L1 1575.42 L2 1227.60 L5 1176.45	L1 1602.00 L2 1246.00 L3 1202.025	B1 1561.098 B2 1207.14 B3 1268.52	E1 1575.42 E5a 1176.45 E5b 1207.14 E6 1278.75	L1 1575.42 L2 1227.60 L5 1176.45 E6 1278.75	L5 1176.45 S 2492.028

SPS: Standard Positioning Service; PPS: Precise Positioning Service; OS: Open Service; AS: Authorized Service; WADS: Wide Area Differential Service; SMS: Short Message Service; CS: Commercial Service; PRS: Public Regulated Service; GCS: GPS Complementary Service; GAS: GPS Augmentation Service; EWS: Early Warning Service; MCS: Message Communications Service; PS: Precision Service; RS: Restricted Service

the launch of the Block IIF satellites (beginning in May 2010) will be interoperable with that of Galileo, QZSS, and IRNSS/NavIC.

GLONASS

The former Soviet Union developed the *Global'naya Navigatsionnaya Sputnikovaya Sistema* or GLONASS (Chap. 8). The first GLONASS satellite was launched on 12 October 1982. By early 1996, a fully operational constellation of 24 satellites was in orbit. Unfortunately, the full constellation was short lived. Russia's economic difficulties following the dismantling of the Soviet Union hurt GLONASS. By 2002 the constellation had dropped to as few as seven satellites, with only six available during maintenance operations. With support of the Russian government, GLONASS was reborn, and on 8 December 2011, full operational capability (FOC) was again achieved and has been subsequently maintained.

The GLONASS satellites are categorized into three different generations: first generation GLONASS I/II (started in 1982), second generation GLONASS-M

(started in 2003), and third generation GLONASS-K (started in 2011). All GLONASS satellites launched since December 2005 have been GLONASS-M satellites with the exception of two GLONASS-K1 satellites, launched on 26 February 2011 and 30 November 2014. GLONASS uses frequency division multiple access (FDMA) for its signals. Originally, the system transmitted the signals within two bands: L1, 1602–1615.5 MHz, and L2, 1246–1256.5 MHz, at frequencies spaced by 0.5625 MHz at L1 and by 0.4375 MHz at L2. GLONASS-K satellites include, for the first time, code division multiple access (CDMA) signals accompanying the legacy FDMA signals. GLONASS-K1 as well as the latest GLONASS-M satellites transmit a CDMA signal on a new L3 frequency (1202.025 MHz).

Galileo

The Galileo system (Chap. 9) is a joint initiative of the European Commission (EC) and the European Space Agency (ESA). The first two In-orbit Validation (IOV) satellites were launched on 21 October 2011, and the

third and fourth **IOV** satellites were launched on 12 October 2012. The first two full-operational-capability satellites, were launched on 22 August 2014, into wrong orbits due to an upper rocket stage anomaly. Up to the end of 2015, six further satellites have been launched. In all, when the constellation is fully developed, there will be 30 Galileo satellites with 24 designated as primary and six spares.

Galileo satellites transmit three levels of service in three frequency bands using CDMA. The Open Service (**OS**) and the Public Regulated Service (PRS) are transmitted in the E1 frequency band centered on 1575.46 MHz (the same as the GPS L1 frequency) and the PRN ranging codes are modulated onto the carrier using binary offset carrier (**BOC**) techniques with each satellite, like GPS, assigned separate codes. The Commercial Service (**CS**) signal and the PRS are transmitted in the E6 frequency band centered on 1278.75 MHz using binary phase-shift keying (**BPSK**) and BOC modulation, respectively. Data and data-less (pilot) signals are transmitted in the E5 frequency band centered on 1191.795 MHz using BOC modulation. Data and pilot signals are also available on E1 and E6. The signals are separated into an E5a and an E5b component and either can be tracked separately or together. The various signals also contain navigation messages supplying the necessary information for acquiring Galileo signals and for determining receiver positions and time.

BeiDou

China fielded a demonstration regional satellite-based navigation system known as BeiDou (Chinese for the *Big Dipper* asterism and pronounced *bay-dough*) following a program of research and development that began in 1980 (Chap. 10). The initial constellation of three GEO satellites was completed in 2003. A fourth GEO satellite was launched in 2007. The initial regional BeiDou system (BeiDou-1) has been replaced by a global system known as BeiDou-2 (or simply BeiDou and formerly known as Compass). The BeiDou Navigation Satellite System (BDS) as it is officially now known will eventually include five GEO satellites, 27 MEO satellites, and five IGSO satellites. BeiDou-2 was declared operational for use in China and surrounding areas on 27 December 2011. FOC for this area was declared on 27 December 2012. The system will provide global coverage by 2020. As of February 2016, 21 BeiDou satellites have been launched. Some of the BeiDou satellites launched from 30 March 2015 onward are a new version termed BeiDou Phase 3 or simply BeiDou-3.

The satellites transmit two levels of service, an open service and an authorized service primarily for the Chinese government and military using three fre-

quency bands. The bands and the central frequencies for the satellites now in use, the BeiDou-2 satellites, are B1 at 1561.098 MHz, B2 at 1207.14 MHz, and B3 at 1268.52 MHz. BeiDou-3 will transmit modernized signals in the L1/E1 and L5/E5 bands as well as the BeiDou B3 band. For compatibility it is also foreseen that they will transmit the B1 open service signal of the BeiDou-2 system.

1.5.2 Regional Navigation Satellite Systems

QZSS

The Quasi-Zenith Satellite System (QZSS; Chap. 11) will use multiple satellites in inclined orbits, placed so that one satellite always appears near zenith above Japan, well known for its high-rise cities where the signals from GPS satellites can be easily blocked. The design provides high-accuracy satellite positioning service covering almost all of the country, including urban canyons and mountainous terrain. The IGSO satellites will be supplemented with one GEO satellite. The start of full QZSS service is planned for 2018. QZSS Phase One is validating technological enhancement of GPS availability, performance, and application using the first QZSS satellite, Michibiki. Michibiki was launched on 11 September 2010 and is in full operation. Phase Two will demonstrate full system capability using at least three QZSS satellites, including Michibiki. Future plans call for a seven-satellite constellation.

The satellites will generate and transmit their own signals, compatible with modernized GPS signals. QZSS also transmits GPS corrections and availability data, the L1-SAIF (Submeter-class Augmentation with Integrity Function) signal, and so is also considered as a satellite-based augmentation system satellite. Altogether, Michibiki transmits six signals with structures similar to and compatible with GPS and Galileo signals: L1-C/A (1575.42 MHz), L1C (1575.42 MHz); L2C (1227.6 MHz), L5 (1176.45 MHz), L1-SAIF (1575.42 MHz), and LEX (L-band Experiment, 1278.75 MHz) a QZSS experimental signal for a high precision (3 cm level) service, sharing the frequency of the Galileo E6 signal.

IRNSS/NavIC

The Indian government has developed the Indian Regional Navigation Satellite System (IRNSS) as an independent system serving India and the surrounding area (Chap. 11). In April 2016, IRNSS was renamed to NavIC, a Hindi word for sailor or navigator as well as an acronym for *Navigation with Indian Constellation*. The area of coverage is from 30° south to 50° north in latitude and 30° east to 130° east in longi-

Table 1.3 An overview of the SBASs

System	WAAS	SDCM	EGNOS	MSAS	GAGAN
Orbit	GEO	GEO	GEO	GEO	GEO
Nominal number of satellites	3	3	4	1	3
Longitudes	133° W, 107° W, 98° W	16° W, 95° E, 167° E	15.5° W, 5° E, 25° E, 31.5° E	145° E	55° E, 83° E, 93.5° E
Date of being operational	July 2003	–	October 2009	September 2007	February 2014
Origin	USA	Russia	Europe	Japan	India
Service area	CONUS, Alaska, Canada, Mexico	Russia	Europe	Japan	India
Frequency (MHz)	L1 1575.42 L5 1176.45	L1 1575.42	L1 1575.42 L5 1176.45	L1 1575.42	L1 1575.42 L5 1176.45

tude. IRNSS provide two types of service: the SPS, which is an open service for all users, and the Restricted Service (RS), which is an encrypted service available only to authorised users. IRNSS is expected to provide a real-time pseudorange-based position accuracy of better than 20 m in the primary service area. The IRNSS constellation consists of three GEO satellites as well as two pairs of IGSO satellites.

The first satellite in the constellation, an IGSO satellite, IRNSS-1A, was launched on 1 July 2013. The second IGSO satellite, IRNSS-1B, was launched on 4 April 2014. The first GEO satellite, IRNSS-1C, was launched on 15 October 2014. IRNSS-1D, the third IGSO satellite, was launched on 28 March 2015. The constellation was completed by launches of IRNSS-1E, 1F, and 1G in 2016. The satellites transmit navigation signals at 1176.45 and 2492.028 MHz in the L- and S-bands respectively. The SPS and RS are transmitted on both frequencies. The SPS uses BPSK modulation while the RS uses BOC modulation with data and pilot channels.

1.5.3 Satellite-Based Augmentation Systems

In addition to the GNSS/RNSSs, there are also SBASs, which use geostationary communications satellites to provide differential correction data and integrity information to GNSS users in real time using a *bent path* from a ground station through the satellite to a user's equipment. The systems use a state-space-domain approach in which corrections for GNSS satellite orbit and clock data along with ionospheric propagation delays are provided.

Currently, four SBASs are in full operation: the FAA's WAAS, the European Geostationary Navigation Overlay Service (EGNOS), Japan's Multifunctional Transport Satellite (MTSAT) Satellite-based Augmentation System (MSAS), and India's GPS-aided GEO Augmented Navigation System (GAGAN). In addition, Japan's Quasi-Zenith Satellite System has an augmentation component as already mentioned. Russia's System for Differential Correction and Monitoring (SDCM) is currently in development. See Table 1.3 for an overview and Chap. 12 for a detailed description of SBASs.

1.6 GNSS for Science and Society at Large

GNSS is used for many types of applications, covering the mass market, professional and safety-critical applications as well as a whole range of scientific applications (Chaps. 29, 30, and 32). There are therefore literally hundreds of applications of GNSS, from the everyday to the exotic. And with the advent of next generation GNSSs, many more applications are expected to emerge.

According to a recent market study [1.26], the global GNSS market is expected to grow from roughly 50B Euros in 2013 to more than 100B Euros in 2023, when considering only the core revenue, i.e., the value of the chipsets sold. Enabled revenues covering the entire end-user equipment are projected to even grow from 200B Euros to almost 300B Euros in the same period. As illustrated in Fig. 1.10, the GNSS market is clearly

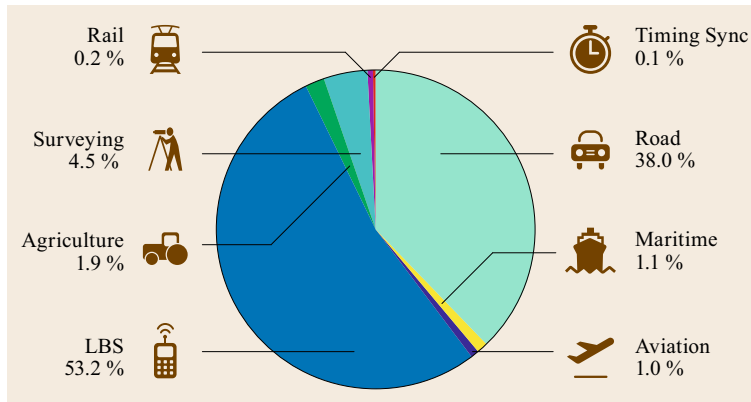


Fig. 1.10 Distribution of cumulative global revenue from GNSS chipset sales projected for the 2013–2023 period (after [1.26], courtesy of European GNSS Agency)

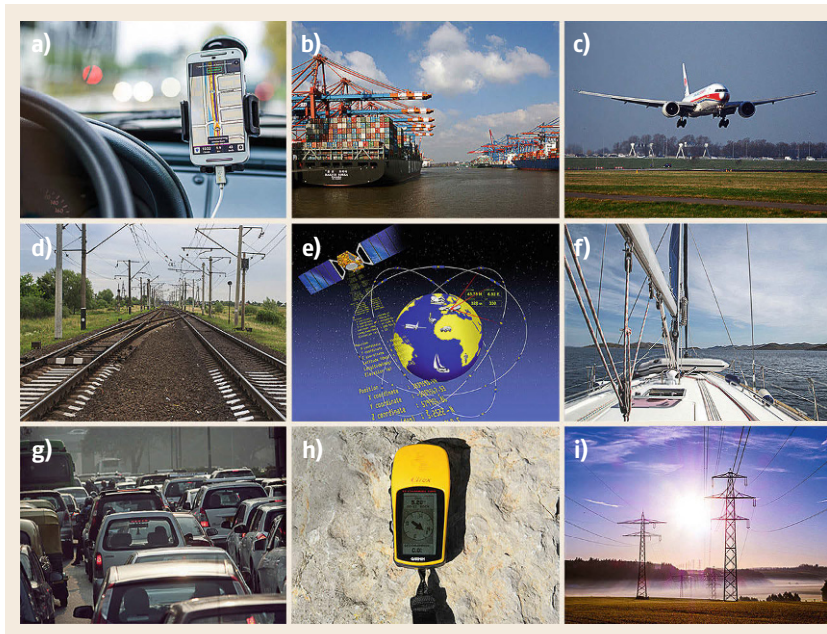


Fig. 1.11a–i Examples of everyday GNSS applications; from car navigation, to the landing of aircraft, to electrical power grid maintenance. (Photos courtesy of [pixabay.com](https://www.pixabay.com) (a–d), (f–i) and ESA, J. Huart (e))

dominated by personal navigation devices (covered by location based services, LBS) and in-car navigation systems, which constitute more than 90% of the global core revenue. High-precision and specialized GNSS equipment for surveying and agriculture as well as maritime and aviation use, in contrast, contributes less than 10% of the overall chipset sales.

By far the most common use of GNSS is for navigation, which includes navigation for people who are hiking and geocaching (a treasure-hunting game); navigation of cars and other vehicles; ocean navigation for marine vessels and channel dredging; and aircraft control, as in approach and landing at airports (Fig. 1.11). Some of these applications, requiring higher accuracy, involve the use of carrier-phase measurements. GNSS is also used for tracking of people, vehicles, vessels,

aircraft, and assets, where GNSS-determined positions are reported via a supplementary communication channel such as that provided by a mobile phone.

One of the earliest high-precision applications of GNSS was in surveying and geodesy (Chaps. 35 and 36) (Fig. 1.12). Through the use of carrier-phase measurements, accurate coordinates of lot boundaries and geodetic markers can be established. Later on these carrier-phase-based techniques found their way into machine control, attitude determination, and precision agriculture [1.27]. Because GNSS also provides precise time (Chap. 41), it is also used to synchronize timing systems worldwide, permitting very accurate time-tagging of financial trades, for example. GNSS timing is widely used in the telecommunication industry including synchronization of mobile phone networks. It is also



Fig. 1.12a–i Examples of high-precision GNSS applications; from surveying and geodesy to machine guidance and precision agriculture (Photos courtesy of Position Partners ((a), left); Leica Geosystems ((a), right); M. Gottlieb, University NAVSTAR Consortium (UNAVCO) (b); pixabay.com (c,f) TU Delft (d); B. Morris (e); Trimble (g); V. Janssen (h); Deere & Co. (i))

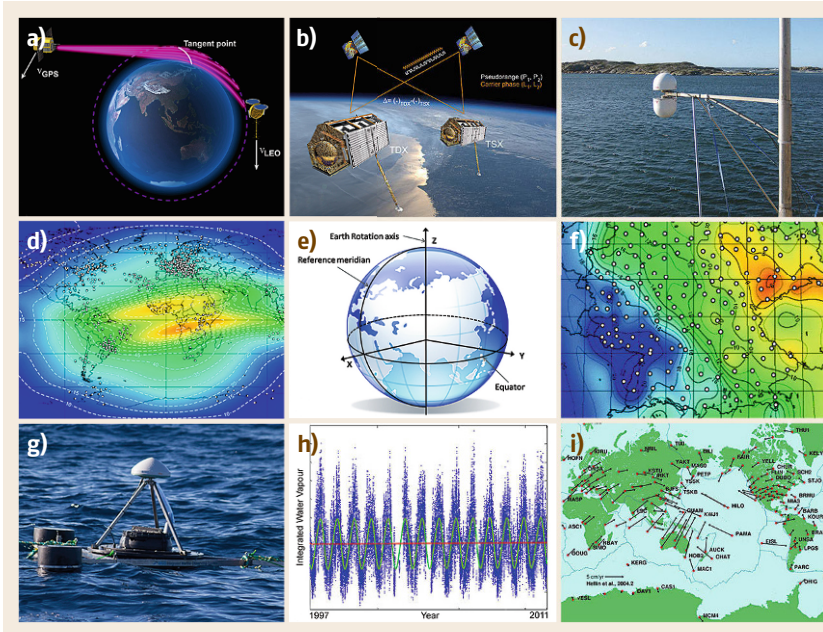


Fig. 1.13a–i Examples of scientific GNSS applications; from atmospheric sensing and reference frame studies to tectonic monitoring (Photos courtesy of UCAR 2007 (a); P. Kuss, DLR, NASA (b); J. Löfgren (c); N. Jakowski, DLR (d); J. Legrand & C. Bruyninx, ROB (e); G. Dick, GFZ (f); IMOS (g); G. Elgered, Chalmers (h); NASA, JPL-Caltech (i))

used for electrical power grids [1.28] to synchronize the phase of alternating current and for power-line fault isolation.

GNSSs are also important for Earth system studies and global environmental Earth observation (Fig. 1.13). Advanced GNSS receivers are operated continuously or periodically at geophysically interesting sites (Chap. 37). Long-term tectonic plate motion is mea-

sured using GNSS [1.29] and networks of receivers are used to assess Earth surface activities for monitoring landslide and volcano activity or the study of land uplift following the last ice age. By monitoring crustal motions with extensive GNSS networks of continuously operating tracking stations, researchers have the long-term goal of being able to make accurate earthquake predictions, and thereby save lives.

GNSSs are also an invaluable instrument for atmospheric sensing (Chaps. 38 and 39). As GNSS signals are affected by their passage through the ionosphere and the lower atmosphere, an appropriate analysis of the received signals can be used to map the ionosphere's variable electron content [1.30] and the amount of water vapor in the troposphere [1.31]. Several national weather services use GNSS-determined water-vapor measurements to improve their weather forecasts. And GNSSs also help us to understand the processes taking place in the ionosphere. This is important as bad weather in the ionosphere may severely disturb our communication, navigation, and power systems. By looking at the effects that earthquakes and tsunamis have on ionospheric electron density [1.32, 33], GNSS ionospheric sensing may become an important component in tsunami warning systems as well.

During the last few years, GNSS data has also proven its potential for climate monitoring. Several

studies have demonstrated the potential of both ground-based and radio occultation GNSS data for the accurate measurement of atmospheric water-vapor content and temperature structure [1.34, 35]. Determination of global temperature trends are important when it comes to the question of global warming. And since GNSS also provides an accurate positioning tool for investigating how ice masses move and change, it also contributes to gaining a better understanding of the linkage between global warming and the melting of our glaciers and the shrinkage of the Arctic ice cover.

Acknowledgments. Some of the material in this chapter stems from the authors lectures on GNSS over the years and the beneficial input of past and present students, research associates, and other colleagues. Some of it is also drawn from the first author's long running Innovation column in GPS World magazine.

References

- 1.1 J.R. Vetter: Fifty years of orbit determination: Development of modern astrodynamics methods, Johns Hopkins APL Tech. Dig. **27**(3), 239–252 (2007)
- 1.2 T.A. Stansell: The Navy Navigation Satellite System: Description and status, Navigation **15**(3), 229–243 (1968)
- 1.3 R.J. Danchik: An overview of Transit development, Johns Hopkins APL Tech. Dig. **19**(1), 18–26 (1998)
- 1.4 P. Daly, G.E. Perry: Recent developments with the Soviet Union's VHF satellite navigation system, Space Commun. Broadcast. **4**, 51–61 (1986)
- 1.5 P. Daly, G.E. Perry: Update on the behaviour of the Soviet Union's VHF satellite navigation system, Space Commun. Broadcast. **5**, 379–384 (1987)
- 1.6 G. Seeber: *Satellite Geodesy: Foundations, Methods and Applications* (Walter de Gruyter, Berlin 2003)
- 1.7 K. Kovach: New user equivalent range error (UERE) budget for the modernized Navstar Global Positioning System (GPS), Proc. ION NTM, Anaheim (2000) pp. 550–573
- 1.8 Global Positioning System Standard Positioning Service Performance Standard (US Department of Defense, Washington DC 2008)
- 1.9 O. Montenbruck, P. Steigenberger, A. Hauschild: Broadcast versus precise ephemerides: A multi-GNSS perspective, GPS Solutions **19**(2), 321–333 (2015)
- 1.10 R. Prieto-Cerdeira, R. Orus-Peres, E. Breeuwer, R. Lucas-Rodríguez, M. Falcone: The European way: Performance of the Galileo single-frequency ionospheric correction during in-orbit validation, GPS World **25**(6), 53–58 (2014)
- 1.11 D. Milbert: Dilution of precision revisited, Navigation **55**(1), 67–81 (2008)
- 1.12 S. Whitney: Global Positioning System status, Proc. ION GNSS+, Tampa (2015) pp. 1193–1206
- 1.13 Minimum Operational Performance Standards for Global Positioning/Wide Area Augmentation System Airborne Equipment (RTCA, Washington DC 2006)
- 1.14 R. Leandro, M. Santos, R.B. Langley: UNB neutral atmosphere models: Development and performance, Proc. ION NTM 2006, Monterey (ION, Virginia 2006) pp. 564–573
- 1.15 R.F. Leandro, R.B. Langley, M.C. Santos: UNB3m_pack: A neutral atmosphere delay package for radiometric space techniques, GPS Solutions **12**(1), 65–70 (2008)
- 1.16 A. Komjathy: Global Ionospheric Total Electron Content Mapping Using the Global Positioning System, Ph.D. Thesis (Univ. New Brunswick, Fredericton 1997)
- 1.17 J.A. Klobuchar: Ionospheric time-delay algorithm for single-frequency GPS users, IEEE Trans. Aerosp. Electron. Sys. **23**(3), 325–331 (1987)
- 1.18 European GNSS (Galileo) Open Service Ionospheric Correction Algorithm for Galileo Single Frequency Users, Iss. 1.2 (European Commission, 2016)
- 1.19 S. Banville, R.B. Langley: Instantaneous cycle-slip correction for real-time PPP applications, Navigation **57**(4), 325–334 (2010)
- 1.20 P.J.G. Teunissen: The Least-squares Ambiguity Decorrelation Adjustment: A method for fast GPS integer ambiguity estimation, J. Geod. **70**(1), 65–82 (1995)
- 1.21 RTCM Standard 10403.2 Differential GNSS Services, Version 3 with Amendment 2 (RTCM, Arlington 2013)
- 1.22 C. Haslett: *Essentials of Radio Wave Propagation* (Cambridge Univ. Press, Cambridge 2008)
- 1.23 A.W. Doerry: *Earth Curvature and Atmospheric Refraction Effects on Radar Signal Propagation* (Sandia National Laboratories, Albuquerque NM 2013), Sandia Report SAND2012-10690

- 1.24 G. Weber, D. Dettmering, H. Gebhard, R. Kalafus: Networked transport of RTCM via internet protocol (Ntrip) – IP-streaming for real-time GNSS applications, Proc. ION GPS, Long Beach (ION, Virginia 2005) pp. 2243–2247
- 1.25 J.G. Walker: Satellite constellations, J. Br. Interplanet. Soc. **37**, 559–572 (1984)
- 1.26 European GNSS Agency: *GNSS Market Report*, 4th edn. (Publications Office of the European Union, Luxembourg 2015)
- 1.27 J.V. Stafford: Implementing precision agriculture in the 21st century, J. Agric. Eng. Res. **76**(3), 267–275 (2000)
- 1.28 A. Carta, N. Locci, C. Muscas, S. Sulis: A flexible GPS-based system for synchronized phasor measurement in electric distribution networks, IEEE Trans. Instrum. Meas. **57**(11), 2450–2456 (2008)
- 1.29 K.M. Larson, J.T. Freymueller, S. Philipsen: Global plate velocities from the Global Positioning System, J. Geophys. Res. Solid Earth **102**(B5), 9961–9981 (1997)
- 1.30 M. Hernández-Pajares, J.M. Juan, J. Sanz: New approaches in global ionospheric determination using ground GPS data, J. Atmos. Sol. –Terr. Phys. **61**(16), 1237–1247 (1999)
- 1.31 M. Bevis, S. Chiswell, T.A. Herring, R.A. Anthes, C. Rocken, R.H. Ware: GPS meteorology: Mapping zenith wet delays onto precipitable water, J. Appl. Meteorol. **33**(3), 379–386 (1994)
- 1.32 E. Calais, J.B. Minster: GPS detection of ionospheric perturbations following the January 17, 1994, Northridge earthquake, Geophys. Res. Letts. **22**(9), 1045–1048 (1995)
- 1.33 A. Komjathy, Y.-M. Yang, X. Meng, O. Verkhoglyadova, A.J. Mannucci, R.B. Langley: Review and perspectives: Understanding natural hazards-generated ionospheric perturbations using GPS measurements and coupled modeling, Radio Sci. **51**(7), 951–961 (2016)
- 1.34 T. Nilsson, G. Elgered: Long-term trends in the atmospheric water vapour content estimated from ground-based GPS data, J. Geophys. Res. **113**(D19101), 1–12 (2008)
- 1.35 R.A. Anthes: Exploring earth's atmosphere with radio occultation: Contributions to weather, climate and space weather, Atmos. Meas. Tech. **4**, 1077–1103 (2011)

2. Time and Reference Systems

Christopher Jekeli, Oliver Montenbruck

Geodesy is the science of the measurement and mapping of the Earth's surface, and in this context it is also the science that defines and realizes coordinates and associated coordinate systems. Geodesy thus is the foundation for all applications of global navigation satellite system (GNSS). This chapter presents the reference systems needed to describe coordinates of points on the Earth's surface or in near space and to relate coordinate systems among each other, as well as to some *absolute* system, visually, a celestial system. The topic is primarily one of geometry, but the geodynamics of the Earth as a rotating body in the solar system plays a fundamental role in defining and transforming coordinate systems. Therefore, also the fourth coordinate, time, is critical not only as the independent variable in the dynamical theories, but also as a parameter in modern geodetic measurement systems. Instead of expounding the theory of geodynamics and celestial mechanics, it is sufficient for the purpose of this chapter to describe the corresponding phenomena, textually, analytically and illustratively, in order to give a sense of the scope of the tasks involved in providing accurate coordinate reference systems not just to geodesists, but to all geoscientists.

2.1	Time	25
2.1.1	Dynamic Time	26
2.1.2	Atomic Time Scales	27
2.1.3	Sidereal and Universal Time, Earth Rotation	27
2.1.4	GNSS System Times	30
2.2	Spatial Reference Systems	31
2.2.1	Coordinate Systems	31
2.2.2	Reference Systems and Frames	34
2.3	Terrestrial Reference System	34
2.3.1	Traditional Geodetic Datums	34
2.3.2	Global Reference System	36
2.3.3	Terrestrial Reference Systems for GNSS Users	39
2.3.4	Frame Transformations	40
2.3.5	Earth Tides	42
2.4	Celestial Reference System	44
2.5	Transformations Between ICRF and ITRF	46
2.5.1	Orientation of the Earth in Space	46
2.5.2	New Conventions	50
2.5.3	Polar Motion	52
2.5.4	Transformations	54
2.6	Perspectives	55
	References	56

2.1 Time

Everyone experiences time, but when pressed no one can explain exactly what it is. Mathematically it can be defined as a coordinate in a fourth dimension (as did Einstein), or more traditionally, it is the independent variable in our theories of motion. Indeed, the only reason that we perceive time is that things change. We have relatively easy access to *units* of time because many of the changes that we observe are periodic. If the changing phenomenon varies with uniform period, then the associated *time scale* is uniform. Clearly, a desirable property of a description and realization of time is that its scale should be uniform at least in the local frame. However, very few observed dynamical systems have rigorously uniform time units. In the past, Earth's

rotation provided the most suitable and evident phenomenon to represent the time scale, with the unit being a (solar) day [2.1]. It has been recognized for a long time, however, that Earth's rotation is not uniform (it is varying at many different scales: daily, bi-weekly, monthly, etc., and even slowing down over geologic time [2.2, p. 607]). In addition to scale or units, an origin must be defined for a time system, that is, a zero-point, or an epoch, at which a value of time is specified. Finally, whatever system of time is defined, it should be accessible and, thereby, realizable, thus creating a time *frame*. This distinction between a system and a frame is explained in greater detail with respect to spatial coordinates in Sect. 2.2.2.

Prior to 1960, a second of time was defined as $1/86\,400$ of a mean solar day (Sect. 2.1.3). Today (since 1960), a fundamental time scale is defined by the natural oscillation of the cesium atom and all time systems can be referred or transformed to this scale. Specifically, the SI (*Système International*) second is defined as follows [2.3, 4]:

The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium 133 atom.

This definition has been refined to specify that the atom should be at rest (i. e., at temperature 0 K) and at mean sea level, thus independent of ambient radiation effects and relativistic gravitational changes. Corrections are applied to actual measurements to comply with these requirements. The value of the SI second was set to the previously (in 1956) adopted value of a second of *ephemeris time* (ET) (Sect. 2.1.1), defined as $1/31\,556\,925.9747$ of a mean tropical (solar) year, being computed for the epoch, 1 January 1900, on the basis of Newcomb's theory of motion of the Earth around the Sun [2.5].

Although the SI second now defines the fundamental time unit, one still distinguishes between systems of time that have different origins and even different scales depending on the application. Dynamic time is the independent variable in the most complete theory of the dynamics of the solar system. It is uniform by definition. Mean solar time, or universal time (UT), is the time scale based on Earth's rotation with respect to the Sun and is used for general civilian time keeping. Finally, sidereal time is defined by Earth's rotation with respect to the celestial sphere. Within this section, the various types of dynamic, atomic, and sidereal time scales are described in further detail.

2.1.1 Dynamic Time

Newtonian (ephemeris time) and relativistic (barycentric and terrestrial time (TT), etc.) concepts of *dynamic time* generally refer to the time variable in the equations of motion describing the dynamical behavior of the massive bodies of our solar system. As such, with respect to the theory of general relativity, the dynamic time scale refers to a coordinate system and thus represent a coordinate time (Chap. 5). Common choices include the *barycentric* reference system (origin at the center of mass of the solar system) or the *geocentric* reference system. The corresponding time scales are thus designated as *barycentric coordinate time* (TCB) and *geocentric coordinate time* (TCG). Note that acronyms for time systems generally follow

the corresponding French names, for example, *temps-coordonnée barycentrique* for Barycentric Coordinate Time. Dynamic time defined in this way is the fourth coordinate and transforms according to the theory of general relativity as the fourth coordinate from one point in space–time to another.

On the other hand, dynamic time has also been defined as a *proper time*, the time associated with the frame of the observer that a uniformly running clock would keep and that describes observed motions in that frame. Depending on the frame of the observer, it is designated, for example, *terrestrial dynamic time* (TDT), or *barycentric dynamic time* (TDB). In 1991, the International Astronomical Union (IAU) renamed TDT simply *terrestrial time*, referring to proper time at the geoid (approximately mean sea level). However, in 2000 the IAU further recommended, due to uncertainties in the realization of the geoid, that TT be redefined as differing from TCG by a constant, specified rate. Its relation to a proper time then more precisely depends on the location and velocity of the observer's clock in the ambient gravitational field. Mathematical connections to the coordinate times, TCB and TCG, and to TDB may be found in Chap. 5 of this Handbook as well as in [2.6, Chap. 10]. The realization of TT is atomic time (Sect. 2.1.2), that is, its scale is the SI second. For calculations of Earth orientation (Sect. 2.5.1), the difference between TT and TDB is usually neglected.

Prior to 1977, the dynamic time was called *ephemeris time*. ET was based on the time variable in the theory of motion of the Sun relative to the Earth – Newcomb's ephemeris of the Sun. This theory suffered from the omission of relativistic theory, the dependence on adopted astronomical constants that, in fact, show a time dependency (such as the *constant* of aberration). It also omitted the effects of other planets on Earth's orbit. The new dynamic time described above was constrained to be consistent with ET at their boundary; specifically,

$$TT = ET \text{ at } 1977 \text{ January } 1.0003725 \\ (1^d 0^h 00^m 32.184^s, \text{ exactly}). \quad (2.1)$$

The extra fraction in this epoch was included since this would make the point of continuity between the systems exactly January 1.0, 1977, in International Atomic Time (TAI) (Sect. 2.1.2).

The basic unit of dynamic time is the *Julian Day*, equal to 86 400 SI seconds, which is close to our usual day based on Earth rotation, but is uniform by definition. The origin of dynamic time, designated by the *Julian date*, or *Julian epoch*, J0.0, is defined to be Greenwich noon, 1 January 4713 BC. Julian days, by convention, start and end when it is noon (dynamic

time) in Greenwich, England, representing midday in the usual meaning of a day starting and ending at midnight. Furthermore, there are exactly 365.25 Julian days in a *Julian year*, or exactly 36 525 Julian days in a *Julian century*. With the origin as given above, the Julian date, J1900.0, corresponds to the Julian day number, 2 415 021.0, being Greenwich noon, January 1, 1900; and the Julian date, J2000.0, corresponds to the Julian day (JD) number 2 451 545.0, being Greenwich noon, January 1, 2000 (Fig. 2.1). Thus, the date with Julian day number 2 451 545.0 is also January 1.5, 2000. Note that January 0.5, 2000 is really Greenwich noon on December 31, 1999 (or December 31.5, 1999). For practical reasons, a *modified Julian day number*

$$\text{MJD} = \text{JD} - 2\,400\,000.5, \quad (2.2)$$

is also defined relative to a new origin, which counts days as starting at midnight in Greenwich.

2.1.2 Atomic Time Scales

Atomic time refers to the time scale defined and realized by the oscillations in energy states of the cesium-133 atom, as defined in the introduction of this section. The SI second thus is the unit that defines the atomic time scale [2.3, 7]. Atomic time was not realized until 1955 with the development of standardized atomic clocks (Chap. 5). From 1958 through 1968, the *Bureau International de l'Heure* (BIH) in Paris maintained the atomic time scale. The origin, or zero point, for atomic time has been chosen officially as $0^{\text{h}} 0^{\text{m}} 0^{\text{s}}$, January 1, 1958.

International Atomic Time was officially introduced in January 1972. It was determined and subsequently defined that on $0^{\text{h}} 0^{\text{m}} 0^{\text{s}}$, January 1, 1977 (TAI), the ET epoch was $0^{\text{h}} 0^{\text{m}} 32.184^{\text{s}}$, January 1, 1977 (ET); thus, in accord with (2.1),

$$\text{TAI} = \text{TT} - 32.184 \text{ s}. \quad (2.3)$$

TAI is realized today by the *Bureau International des Poids et Mesures* (BIPM), which combines data from over 400 high-precision atomic clocks around the world in order to maintain the SI-second scale as accurately as possible. TAI is published and accessible as a correction to each time-center clock, but rather in terms of *coordinated universal time* (UTC, Sect. 2.1.3), which is civilian atomic time adjusted to be close to a time scale based on Earth's rotation.

In the United States, the official atomic time clocks are maintained by the US Naval Observatory (USNO) in Washington, DC, and by the National Institute of Standards and Technology (NIST) in Boulder, CO, USA. Within each such center several cesium beam clocks are running simultaneously and averaged. Other centers participating in the realization of TAI include observatories in Paris, Greenwich, Moscow, Tokyo, Ottawa, Wettzell, Beijing, and Sydney, among over 70 others. The comparison and amalgamation of the clocks of participating centers around the world are accomplished by LORAN-C, satellite transfers (GNSS playing the major role; Chap. 41), and actual clock visits. Time offsets of individual laboratories and their uncertainties are reported in the monthly issues of the BIPM Circular T [2.8]. Worldwide synchronization for many of the national laboratories is at the level of a few tens of nanoseconds or better [2.9]. Since atomic time is computed from many clocks, it is also known as a *paper clock* or a *statistical clock*.

2.1.3 Sidereal and Universal Time, Earth Rotation

Sidereal time represents the rotation of the Earth with respect to the celestial sphere and reflects the actual rotation rate of the Earth, plus effects due to the small motion of the spin axis relative to space (precession and nutation, Sect. 2.5.1). It is the angle on the equator between a particular terrestrial meridian and the *vernal equinox*, Υ , which is the point on the celestial sphere where the Sun crosses the equator in Spring

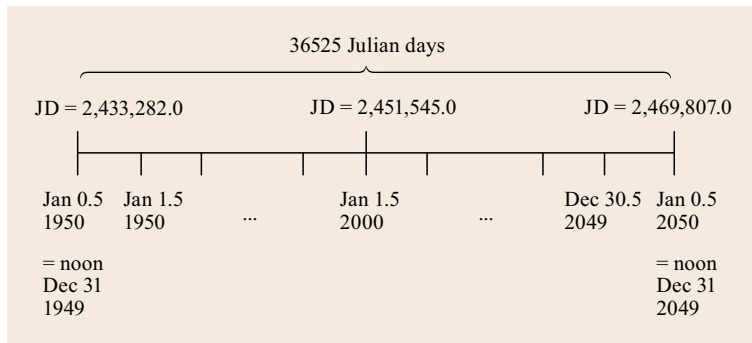


Fig. 2.1 Julian Day numbers and their relation to our current conventional calendar

as viewed by the Northern Hemisphere of the Earth. Inasmuch as the equator has the same dynamics as the spin axis, one distinguishes between apparent (or, true) and mean sidereal time, the latter having the effects of nutation removed. The amplitude of this effect is about $15.8''$, which corresponds to about 1 s of time using the conversion, $15^\circ = 1 \text{ h}$. *Greenwich apparent sidereal time* (GAST) is the angle from the true (or, instantaneous) equinox to the Greenwich meridian (Fig. 2.2).

Due to the precession of the spin axis and thus the vernal equinox on the equator, sidereal time includes a small rotation rate (about $7.1 \cdot 10^{-12} \text{ rad/s}$) that is not due to Earth rotation. For this reason, a new origin point, σ , has been introduced and adopted in the late 1990s that better serves the determination of Earth's rotation rate. This so-called nonrotating origin is also called the celestial intermediate origin (CIO) as explained in Sect. 2.5.2. A new angle, θ , called the Earth rotation angle (ERA), now represents true Earth rotation (Fig. 2.2). The angle $\alpha(\Upsilon) = \theta - \text{GAST}$, also called the equation of origins (EO), today (2015) has a significant value of about $-12'$ due to the accumulated precession since J2000. Expressions for evaluating the EO at arbitrary epochs are provided in [2.6, 10].

UT is the time scale used for general civilian time keeping and is based approximately on the diurnal motion of the Sun. However, the Sun, as viewed by a terrestrial observer does not move uniformly on the celestial sphere. To create a uniform time scale requires the notion of a fictitious, or *mean Sun*, and the corresponding time is known as *mean solar time* (MT). UT is defined as mean solar time on the Greenwich meridian. The basic unit of UT is the *mean solar day*, being the time interval between two consecutive transits of the mean Sun across the meridian. The mean solar day has 24 mean solar hours and 86 400 mean solar seconds.

In comparison to sidereal time, the following approximate relations hold

$$\begin{aligned} 1 \text{ mean solar day} \\ = 24^{\text{h}} 03^{\text{m}} 56.5554^{\text{s}} \text{ in sidereal time,} \end{aligned} \quad (2.4)$$

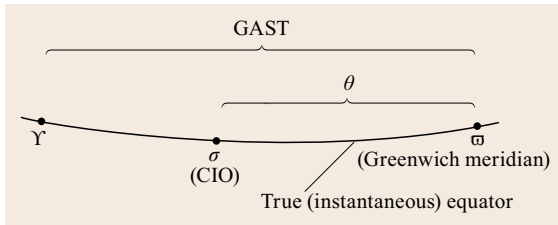


Fig. 2.2 Relationship between GAST and the Earth rotation angle, θ , relative to the true vernal equinox, Υ

$$\begin{aligned} 1 \text{ mean sidereal day} \\ = 23^{\text{h}} 56^{\text{m}} 04.0905^{\text{s}} \text{ in solar time.} \end{aligned} \quad (2.5)$$

A mean solar day is longer than a sidereal day because in order for the Sun to return to the observer's meridian, the Earth must rotate an additional amount due to its orbital advance (Fig. 2.3). Thus, also Earth's rotation rate is *not* equal to $2\pi/86\,400 \text{ rad/s}$ if s is a solar second. Instead, the rate is, according to (2.5),

$$\omega_{\oplus} = 7.292115 \cdot 10^{-5} \text{ rad/s.} \quad (2.6)$$

To determine ω_{\oplus} (and its variations) from measurements by terrestrial observers, one must account for the fact that the observer's reference meridian is associated with a fixed pole, with respect to which the Earth's spin axis moves (polar motion, Sect. 2.5.3). In addition, Earth's rotation is affected by other irregularities of periodic and secular character (such as seasonal effects and the exchange of angular momentum between the Earth and Moon) that are lumped into so-called *length-of-day variations*. Universal time as a scale derived from Earth's rotation has thus been separated into:

- UT0: Universal Time determined from observations with respect to the meridian fixed to the reference pole;
- UT1: Universal Time determined with respect to the meridian attached to the spin axis;
- UT2: Universal Time UT1 corrected for seasonal variations.

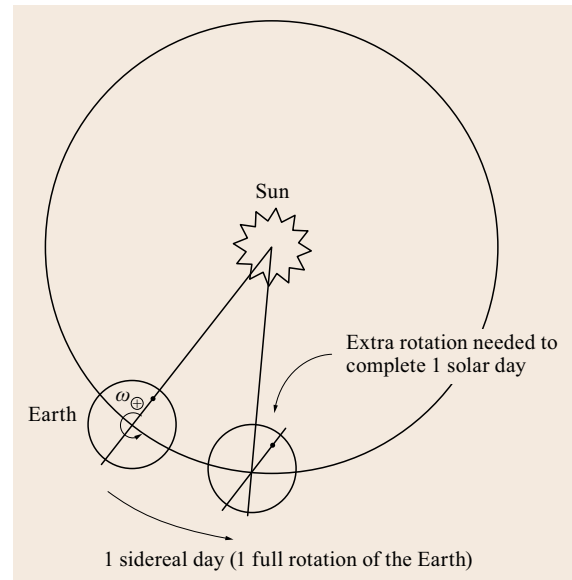


Fig. 2.3 Geometry of sidereal and solar days

UT2 is the best approximation of UT to a uniform time, although it is still affected by small secular variations. However, as a matter of practical utilization it has now been replaced by an atomic time scale (UTC, see below).

In terms of the SI second, the mean solar day is given by

$$1^d (\text{MT}) = 86\,400 \text{ s} - \frac{\Delta\tau}{n}, \quad (2.7)$$

where

$$\Delta\tau = \text{UT1} - \text{TT} \quad (2.8)$$

is the difference over a period of n days between UT1 and dynamic time. The length-of-day variation is the time-derivative of $\Delta\tau$. From observational records over the centuries it has been found that the secular variation in the length of a day (rate of Earth rotation) currently is approximately +1.4 ms per century [2.2, p. 607].

All civilian clocks in the world are now set with respect to an atomic time standard since atomic time is much more uniform than solar time and more easily realized through time transfer by satellite signals. Yet, there is still a desire (particularly, in the astronomic community) that civil time should correspond to solar time; therefore, a new atomic time was defined that approximates UT. This atomic time is called *Coordinated Universal Time* (UTC) and implemented in accord with Recommendation TF.460 of the International Telecommunication Union (ITU) [2.11]:

UTC is the time scale maintained by the BIPM, with assistance from the IERS, which forms the basis of a coordinated dissemination of standard frequencies and time signals. It corresponds exactly in rate with TAI but differs from it by an integral number of seconds. The UTC scale is adjusted by the insertion or deletion of seconds (positive or negative leap seconds) to ensure approximate agreement with UT1.

Initially, UTC was adjusted so that $|\text{UT2} - \text{UTC}| < 0.1 \text{ s}$. As of 1972, the requirement for the correspondence between UTC and UT was relaxed to

$$|\text{UT1} - \text{UTC}| < 0.9 \text{ s}. \quad (2.9)$$

The adjustments, called *leap seconds*, are introduced either January 1 or July 1 of any particular year.

Up to 2015, leap seconds have, on average, been introduced approximately once every 1.5 years (Table 2.1). Following an earlier adjustment in July 2012, the UTC – TAI amounts to –36 s since mid 2015. The

Table 2.1 UTC leap seconds introduced since 1972. The table provides the integer seconds difference between UTC and TAI along with the starting date of applicability (after [2.12])

Since	UTC – TAI (s)	Since	UTC – TAI (s)
1 Jan 1972	–10	1 Jan 1988	–24
1 Jul 1972	–11	1 Jan 1990	–25
1 Jan 1973	–12	1 Jan 1991	–26
1 Jan 1974	–13	1 Jul 1992	–27
1 Jan 1975	–14	1 Jul 1993	–28
1 Jan 1976	–15	1 Jul 1994	–29
1 Jan 1977	–16	1 Jan 1996	–30
1 Jan 1978	–17	1 Jul 1997	–31
1 Jan 1979	–18	1 Jan 1999	–32
1 Jan 1980	–19	1 Jan 2006	–33
1 Jul 1981	–20	1 Jan 2009	–34
1 Jul 1982	–21	1 Jul 2012	–35
1 Jul 1983	–22	1 Jul 2015	–36
1 Jul 1985	–23	1 Jan 2017	–37

history of UTC relative to TAI and other time scales is schematically shown in Fig. 2.4 based on tabulated data of the United States Naval Observatory (USNO) in [2.12]. The decision to introduce new leap seconds is taken by the *International Earth Rotation and Reference Systems Service* (IERS) and announced within the IERS Bulletin C.

The lengthening of a day by about 1.4 ms per century as measured by Earth’s slowing rate of rotation implies that the UT1 clock continues to run more and more behind the TAI clock. It has been determined that the mean solar day today is actually about 86 400.0027 SI seconds long, since the SI second was originally identified with the ET second based on the motion of the mean Sun at Newcomb’s time in the nineteenth century. Indeed, 86 400 SI seconds exactly equaled a mean solar day in 1820, or 1.95 centuries (cy) ago. This disparity between the scales of the defined SI second and the current mean solar day has an accumulative effect that adds, on the average, about $1.4 \text{ ms/d/cy} \times 1.95 \text{ cy}$, or about 1 s to UT1 during the course of a year; hence, the introduction of the leap seconds. The difference, $\text{DUT1} = \text{UT1} - \text{UTC}$, is broadcast along with UTC so that users can determine UT1.

The relationships among the various atomic time scales are illustrated along with dynamic time in Fig. 2.4. There is current debate [2.13–15] about the need to maintain the small difference between UTC and UT1 considering the technical inconveniences and inefficiencies (if not outright difficulties) this imposes on the many modern civilian telecommunications systems and other networks that rely on a precise time scale.

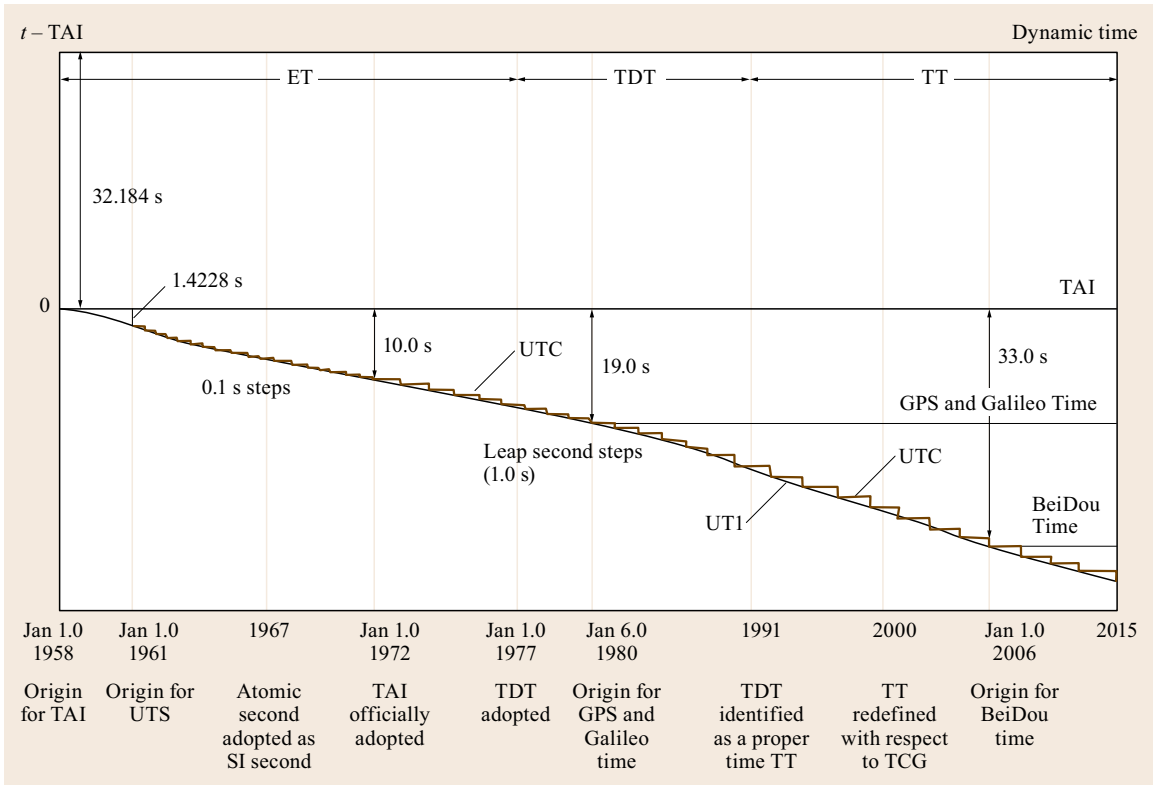


Fig. 2.4 Relationships between atomic time scales and dynamic time (indicated leap seconds are schematic only). For the acronyms, see the text

2.1.4 GNSS System Times

Satellite navigation systems provide user coordinates derived from distance measurements that are based on the propagation time of the transmitted signals. Thus, all these systems rely on very accurate clocks and time standards. To meet the needs of internal time synchronization and dissemination, each GNSS maintains a specific *system time*. The time systems of the four global navigation satellite systems, Global Positioning System (GPS), GLONASS, Galileo, and BeiDou, are all based on the SI second and atomic time similar to TAI. However, they are realized by different clock ensembles and have different origins and offsets with respect to TAI [2.16].

GPS time (GPST) is the system time employed by the United States' Global Positioning System. Since 1990, it is formed as a *composite clock* from atomic clocks within the GPS Control Segment (including both the Master Control Station and the Monitoring Stations) as well as the atomic frequency standards on-board the GPS satellites [2.17, 18]. Each of these clocks contributes to the resulting time scale with a specific

weight based on the observed variance of the respective clock [2.19]. Using common view time transfer, GPS time is steered to deviate by at most $1 \mu\text{s}$ [2.20] from UTC(USNO), that is, the realization of UTC maintained by the United States Naval Observatory. In practice, the GPS–UTC(USNO) offset is much smaller than the specified range and achieves representative values at the level of 20 ns [2.21]. In order to provide GPS users with access to UTC, a forecast value of the offset between both time scales is transmitted as part of the navigation message.

The origin of GPS time, as noted in Fig. 2.4, is January 6.0, 1980 UTC(USNO). However, GPS time is not adjusted by leap seconds to slow down with UT and it is thus permanently offset (late) by a constant amount from TAI

$$t(\text{GPS}) = \text{TAI} - 19 \text{ s} . \tag{2.10}$$

At the same time, it is offset from (ahead of) UTC by varying amounts depending on the number of introduced leap seconds. Note that (2.10) describes only the nominal (integer second) offset between GPS time and

TAI, but neglects additional fractional offsets (typically at the level of tens of nanoseconds) related to different realization of the two time scales.

GLONASS Time (GLST) is the only GNSS time scale that actually follows the ITU recommendation [2.11] to align a disseminated time scale with UTC. Its origin is chosen as January 1.0, 1996 in the UTC(SU) time system, that is, the Russian (formerly Soviet Union, SU) realization of UTC maintained by the Institute of Metrology for Time and Space in Moscow. Besides incorporating leap seconds, GLST is always 3 h ahead of UTC because of the time zone difference between Greenwich and Moscow. Thus,

$$t(\text{GLONASS}) = \text{UTC} + 3 \text{ h} . \quad (2.11)$$

Again, this relation does not account for fractional second offsets resulting from the independent realization of both time scales. GLST is obtained from an ensemble of hydrogen-masers in the GLONASS ground segment and synchronized to UTC(SU) using two-way time transfer with a specified tolerance of 1 μs [2.22]. Following a consolidated effort to improve the alignment of GLST with UTC, the difference of

the two time scales has improved from several hundred ns [2.23] to a few tens of ns as of the second half 2014 [2.24].

Both the *Galileo System Time* (GST, [2.25, 26]) and *BeiDou time* (BDT; [2.27]) exhibit a constant offset from TAI. The origin for Galileo time, for consistency, is defined to be identical to that of GPS Time, but the origin for the BeiDou time system has been chosen as January 1.0, 2006 UTC. Thus,

$$t(\text{Galileo}) = \text{TAI} - 19 \text{ s} , \quad (2.12)$$

$$t(\text{BeiDou}) = \text{TAI} - 33 \text{ s} . \quad (2.13)$$

Both time scales are generated from atomic clocks in the respective control segments and steered to UTC via time transfer and clock comparison with other UTC laboratories. GST is specified to differ by less than 50 ns (2σ) from UTC [2.25, 28] while a maximum offset of 100 ns applies for BeiDou [2.23, 29].

Similar to Galileo, continuous time scales with a fixed -19s offset from TAI are also adopted by the Japanese Quasi-Zenith Satellite System (QZSS) and the Indian Regional Satellite Navigation System (IRNSS/NAVIC).

2.2 Spatial Reference Systems

To establish coordinates of points requires that we set up a coordinate system with origin, orientation, and scale defined in such a way that all users have access to these. Before the establishment of GNSS, the most accessible reference for coordinates from a global perspective was the celestial sphere of stars that were used not only for charting and navigation, but also served as a fundamental system to which other terrestrial coordinate systems could be oriented. Still today, the celestial reference system is used for that purpose and may be thought of as the ultimate in reference systems. At the next level, we define coordinate systems attached to the Earth with various origins (and perhaps different orientations and scale). Thus, there are two fundamental tasks: (1) to establish an external coordinate system of the local universe that presumably remains fixed in the sense of no rotation; and (2) to establish a coordinate system attached to the rotating and orbiting Earth, and in so doing to find the relationship between these two systems.

2.2.1 Coordinate Systems

The Cartesian system of coordinates, x, y, z , is certainly the easiest from a mathematical perspective and

it plays a central role in defining modern reference systems. However, because the Earth is nearly spherical and by extension our geocentric view of the heavens takes on a spherical character, spherical coordinates are essential as many geodetic concepts rely on directions and distances. Indeed, the latitude/longitude concept will always have the most direct appeal for terrestrial applications (surveying, near-surface navigation, positioning, and mapping). Figure 2.5 shows the relationship between the Cartesian coordinates and *spherical coordinates*, comprising latitude, ϕ , longitude, λ , and radius, r , and given by

$$\begin{aligned} x &= r \cos \phi \cos \lambda , \\ y &= r \cos \phi \sin \lambda , \\ z &= r \sin \phi . \end{aligned} \quad (2.14)$$

The inverse relationship is

$$\begin{aligned} \phi &= \tan^{-1} \left(\frac{z}{\sqrt{x^2 + y^2}} \right) , \\ \lambda &= \tan^{-1} \left(\frac{y}{x} \right) , \\ r &= \sqrt{x^2 + y^2 + z^2} . \end{aligned} \quad (2.15)$$

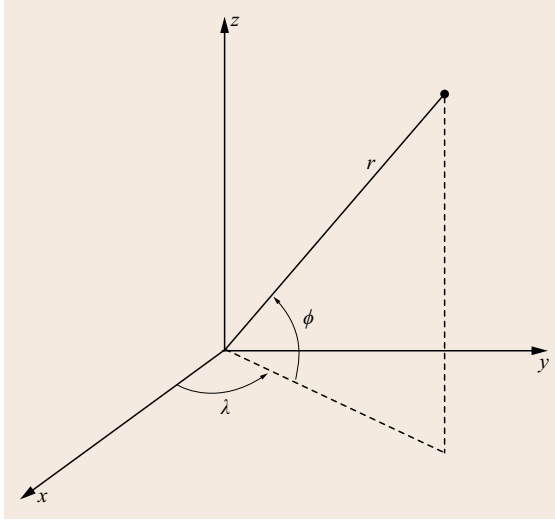


Fig. 2.5 Spherical coordinates

Already by the middle of the eighteenth century, it was established by measurements that the Earth is flattened at the poles and assumes an elliptical shape [2.30], specifically an *ellipsoid* of revolution, defined as the surface generated by rotating an ellipse about its minor axis. It is also known as a *spheroid* (to distinguish it from a tri-axial ellipsoid). Essential parameters of the ellipsoid are its size and shape that may be defined by the semi-major and semi-minor axis lengths, a and b (Fig. 2.6). Other shape parameters include the flattening

$$f = \frac{a-b}{a}, \quad (2.16)$$

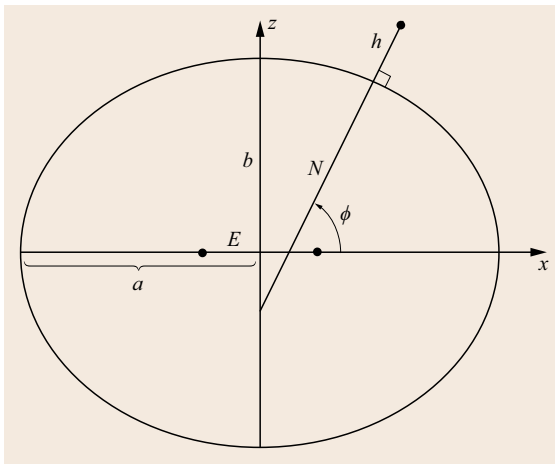


Fig. 2.6 Ellipsoidal geometry and geodetic coordinates. Dots on the x -axis denote focal points of the ellipse, which represents the meridian plane

the first and second eccentricities

$$e^2 = \frac{a^2 - b^2}{a^2} \quad \text{and} \quad e'^2 = \frac{a^2 - b^2}{b^2}, \quad (2.17)$$

as well as the linear eccentricity $E = ae$.

With respect to an ellipsoid with given parameters, the *geodetic coordinates* are defined as illustrated in Fig. 2.6 and include the geodetic latitude, φ , the geodetic longitude, λ (not shown, but identical to the spherical longitude), and the geodetic height, h , along the line that is normal, or perpendicular, to the ellipsoid. The relationship between geodetic coordinates and the global Cartesian coordinates is

$$\begin{aligned} x &= (N + h) \cos \varphi \cos \lambda, \\ y &= (N + h) \cos \varphi \sin \lambda, \\ z &= [N(1 - e^2) + h] \sin \varphi, \end{aligned} \quad (2.18)$$

where

$$N = \frac{a}{\sqrt{1 - e^2 \sin^2 \varphi}}. \quad (2.19)$$

is the radius of curvature of the ellipsoid in the direction perpendicular to the elliptical meridian plane.

An inverse relationship can be formulated for $z \neq 0$ that requires a numerical iteration on the geodetic latitude,

$$\varphi = \tan^{-1} \left[\frac{z}{\sqrt{x^2 + y^2}} \left(1 + \frac{e^2 N \sin \varphi}{z} \right) \right], \quad (2.20)$$

where the initial latitude that assumes the point is on the ellipsoid ($h = 0$),

$$\varphi^{(0)} = \tan^{-1} \left[\frac{z}{\sqrt{x^2 + y^2}} \left(1 + \frac{e^2}{1 - e^2} \right) \right], \quad (2.21)$$

serves to yield convergence to micro-arcsecond accuracy within three iterations for heights less than 20 km. The height then follows from

$$h = \left(\sqrt{x^2 + y^2} \right) \cos \varphi + z \sin \varphi - a \sqrt{1 - e^2 \sin^2 \varphi}, \quad (2.22)$$

and the longitude is given by the second equation of (2.15).

A noniterative relationship is derived by [2.31] based on the solution to a quartic equation; see also [2.32]. The performance and computational efficiency of different analytical and iterative algorithms

for the conversion of Cartesian to geodetic coordinates is, furthermore, compared in [2.33].

A number of ellipsoids have been established on the basis of geodetic measurements, extending historically from surveyed arc lengths along meridians to modern best fits to mean sea level using satellite altimetry. One of the earliest ellipsoids was computed by Airy in 1830, having semi-major axis, $a = 6\,377\,563.396$ m, and flattening, $f = 1/299.324964$. The current internationally adopted ellipsoid is part of the Geodetic Reference System of 1980 (GRS80) and has parameter values given by

$$\begin{aligned} a_{\text{GRS80}} &= 6\,378\,137 \text{ m}, \\ f_{\text{GRS80}} &= \frac{1}{298.257222101}. \end{aligned} \quad (2.23)$$

The equatorial radius was determined from satellite altimetry and the flattening was derived from the second-degree zonal harmonic coefficient (dynamic form factor, J_2) of the Earth's gravitational potential [2.34]. The parameter values of other ellipsoids determined and used in the past may be found in [2.30]. The parameter estimates of the best fitting, or mean Earth ellipsoid (MEE) in the mean tide system are [2.35]

$$\begin{aligned} a_{\text{MEE}} &= 6\,378\,136.72 \pm 0.1 \text{ m}, \\ f_{\text{MEE}} &= \frac{1}{298.25231 \pm 0.00001}. \end{aligned} \quad (2.24)$$

The GRS80 values are constants, while the MEE values are estimates with a standard deviation and do not constitute an accepted reference ellipsoid. When publishing geodetic coordinates, φ, λ, h , it is always important to specify the associated ellipsoid on which they depend.

Local coordinates in the vicinity of a point P are Cartesian with the third axis along the ellipsoid normal

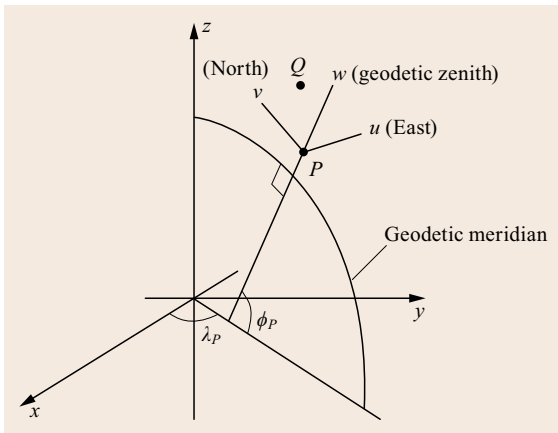


Fig. 2.7 Local Cartesian coordinates, u, v, w

as illustrated in Fig. 2.7. For a right-handed system, the first axis points East and the second North. However, a left-handed system, such as North-East-Up, is also common. Local coordinates $(u, v, w)^T$ of a point Q in a system centered at P are related to the global Cartesian coordinate differences $(\Delta x, \Delta y, \Delta z)^T$ of Q with respect to P according to

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \mathbf{E} \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} \quad (2.25)$$

with

$$\mathbf{E} = \begin{pmatrix} -\sin \lambda & +\cos \lambda & 0 \\ -\sin \varphi \cos \lambda & -\sin \varphi \sin \lambda & +\cos \varphi \\ +\cos \varphi \cos \lambda & +\cos \varphi \sin \lambda & +\sin \varphi \end{pmatrix}. \quad (2.26)$$

Here, the latitude φ and longitude λ refer to the reference point P. The inverse relationship is obtained by premultiplying both sides by the transpose of the rotation matrix since it is orthogonal.

The elevation angle E of Q relative to P and the corresponding azimuth angle A (measured clockwise from North to East) are given by

$$\begin{aligned} \tan A &= \frac{u}{v} \\ \sin E &= \frac{w}{\sqrt{u^2 + v^2 + w^2}}. \end{aligned} \quad (2.27)$$

These formulas relate global Cartesian coordinate differences, as might be obtained by GNSS, to local determinations of angles and distances. If those angles are referenced to the local plumb line, rather than the ellipsoidal normal, one needs to account for this *deflection of the vertical*. For a distance of 1 km and a vertical deflection of $30''$, the effect on the global Cartesian coordinate differences is of the order of a few centimeters or decimeters.

Celestial coordinates refer to the location of objects (e.g., stars) projected onto the *celestial sphere*. By definition, the celestial sphere has no particular radius as the coordinates define only directions. The center of the sphere is defined to be at the origin of a Cartesian coordinate system, and the celestial coordinates are called declination (δ) and right ascension (α), analogous to latitude and longitude. As such, the relationship to Cartesian coordinates is the same as in (2.14) and (2.15) with unit radius ($r = 1$). The origins for declination and right ascension require particular definitions associated with a reference system. This is discussed further in Sect. 2.4.

2.2.2 Reference Systems and Frames

There is an important conceptual difference between a reference system for coordinates and a reference frame that applies throughout the discussion of coordinate systems in geodesy. Loosely recognized in defining and creating geodetic datums in the past, it was formalized by [2.36] (see also [2.37, Chap. 9] and [2.6]):

- A *reference system* is a set of prescriptions and conventions together with the modeling required to define at any time a triad of coordinate axes.
- A *reference frame* realizes the system by means of coordinates of definite points that are accessible directly by occupation or by observation.

A simple example of a reference system is the set of three mutually orthogonal axes that are aligned with the Earth's spin axis, a prime (Greenwich) meridian, and a third direction orthogonal to these two in the right-handed sense. That is, a system defines how the axes are to be established (e.g., orthogonality), what theo-

ries or models are to be used (e.g., what is meant by a spin axis), and what conventions are used (e.g., how the prime meridian is to be chosen). A simple example of a frame is a set of points globally distributed whose coordinates are given mutually consistent numbers in the reference system. That is, a frame is the physical realization of the system defined by actual coordinate values of actual points in space that are accessible to anyone. A frame cannot exist without a system, and a system is of no practical value without a frame.

Although the explicit difference between frame and system was articulated fairly recently in geodesy, the concepts have been embodied in the terminology of a *geodetic datum* that can be traced to the eighteenth century and earlier [2.30]. Indeed, the definition of a datum today refers specifically to the conventions that establish how a coordinate system is attached to the Earth – its origin, its orientation, and its scale. In this sense, the definition of a datum has not changed. The meaning of a datum within the context of frames and systems is explored in more detail in Sect. 2.3.2.

2.3 Terrestrial Reference System

Geodetic control at local, regional, national, and international levels has been revolutionized by the advent of satellite systems, particularly GNSS that provide accurate positioning capability to terrestrial observers at all scales, where, of course, the GPS has had the most significant impact. The terrestrial reference systems and frames for geodetic control have evolved correspondingly over the last few decades. Countries and continents around the world are revising, redefining, and updating their fundamental networks to take advantage of the high accuracy, the ease of establishing and densifying the control, and critically important, the uniformity of the accuracy and the connectivity of the control that can be achieved basically in a global setting.

2.3.1 Traditional Geodetic Datums

The traditional *geodetic datum* was defined somewhat loosely by today's standards as a set of constants and prescriptions that specify a coordinate system for the purpose of geodetic control [2.38]. Because of the fundamental differences in respective measurement techniques, control was divided between horizontal and vertical datums.

Horizontal datums (Fig. 2.8) required the definition of an origin point (a marker on the Earth's surface with defined geodetic latitude and longitude; or, equivalently, a constraint within a network that essentially

fixed the origin), as well as a mapping surface, an ellipsoid with defined parameters. Orientation of the ellipsoid was defined to be parallel to the astronomic system of the celestial sphere (Sect. 2.4). It was realized by accurate measurements of azimuth with respect to celestial north and by accounting for the deflection of the vertical in astronomic determinations of coordi-

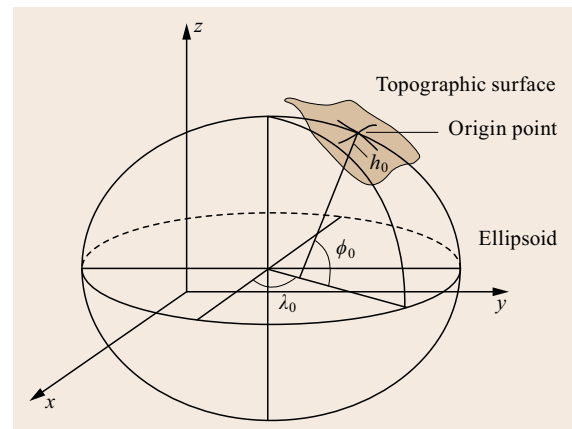


Fig. 2.8 Traditional horizontal geodetic datum. Geodetic surveys on the topographic surface relative to an origin point are reduced to a mapping surface, the ellipsoid, with proper preservation of its orientation relative to an astronomic system

nates. A *vertical datum* (Fig. 2.9) was similarly defined by the height at an origin point and prescriptions for the reference surface through that point and the associated heights relative to the surface.

In the United States, horizontal control was established in the latter half of the nineteenth century for the Eastern United States and advanced with the westward economic expansion to create the *North American Datum of 1927* (NAD27) with origin point at Meades Ranch in the centrally located state of Kansas. In 1983, the horizontal datum was redefined to be geocentric (origin at the now practically accessible center of mass of the Earth by tracking Earth-orbiting satellites), referred to the GRS80 ellipsoid, and readjusted with the inclusion of satellite Doppler observations and other space techniques such as very long baseline interferometry (VLBI [2.39,40]). The new *North American Datum of 1983* (NAD83), already incorporating three-dimensional coordinates, assumed a fully three-dimensional character with each new realization that was adjusted by including continuously operating reference stations (CORS [2.41]). The CORS network is a cooperative endeavor among the US government (National Geodetic Survey) and academic and private institutions that creates precise geodetic control throughout the United States and several worldwide stations using GNSS data. New realizations of NAD83 were adjusted as the CORS network expanded and were designated NAD83(CORS93), NAD83(CORS94), and NAD83(CORS96). Including also additional regional high-accuracy GPS networks that were adjusted to fit the NAD83(CORS96) frame, it became the geometric part of the National Spatial Reference System, designated NAD83(NSRS2007). This was readjusted in 2011, yielding the realization NAD83(2011) with coordinates and their velocities (Sect. 2.3.4) given for the epoch $t_0 = 2010.0$. The reference system definition is currently (2015) in revision to bring the realization closer to the International Terrestrial Reference Frame (ITRF) (Sect. 2.3.2).

The vertical datum in the United States similarly evolved from an adjustment of coast-to-coast leveling networks constrained to zero height at various tide-gauge stations at mean sea level. This *National Geode-*

tic Vertical Datum of 1929 (NGVD29) was replaced in 1988 with a readjustment of existing and new leveling data and a tie to the *International Great Lakes Datum of 1985* (IGLD85) whose origin is a single point on the St. Lawrence River in the province of Québec, resulting in the *North American Vertical Datum of 1988* (NAVD88). Vertical control in the United States and Canada is now undergoing a fundamental redefinition to eliminate continent-wide error trends by defining the reference, not by any particular origin point, but by a model for the Earth's gravity potential. This new *geopotential reference system* already exists for Canada as of 2013, and is scheduled to be in place for the United States by the early 2020s.

Similar progress in geodetic control is occurring in other regions of the world, for example, in Europe and South America, where in some cases progress is more difficult due to the varied and heterogeneous datums established in the pre-satellite era.

While geodetic control is now essentially three-dimensional within a single reference system and frame, such as NAD83(NSRS2007), vertical datums continue to be vitally important since they define a different kind of height, one that is based on gravity potential, rather than pure geometry. The geopotential-based heights are needed for any application in hydrology since they indicate the natural flow of water.

The conversion between ellipsoidal heights, h , obtained from coordinates in the modern geodetic reference system and heights, H , in a vertical datum requires a model for the *geoid undulation*, or *geoid height*, N , defined as the vertical separation between the geoid and the ellipsoid (Fig. 2.10)

$$N = h - H - N_0. \quad (2.28)$$

The *geoid* is the equipotential surface that closely approximates global mean sea level and the geoid undulation is determined from gravity measurements [2.42]. High-degree and high-order spherical harmonic gravitational potential models such as EGM2008 can provide global geoid undulations with an accuracy of 10 cm or better as shown in [2.43]. In addition, a constant offset, N_0 , must be determined between the geoid and the ver-

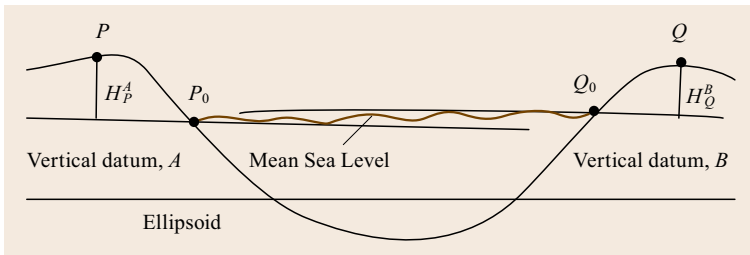


Fig. 2.9 Traditional vertical geodetic datum, representing an equipotential, or level, surface in Earth's gravity field. Since mean sea level is not truly level, different vertical datums tied to mean sea level are not mutually consistent

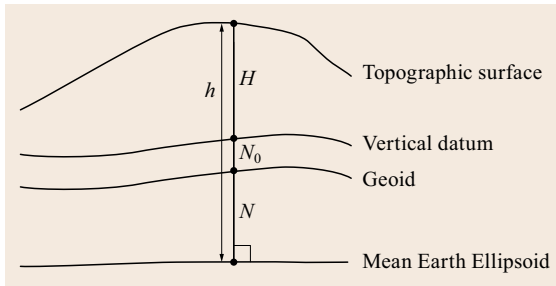


Fig. 2.10 The relationship between heights with respect to a vertical datum and the ellipsoid

tical datum, as well as a possible difference between the best-fitting MEE and the ellipsoid of the reference system. This offset can reach several decimeters in value.

The geoid undulation itself covers a range of roughly ± 100 m with positive peak values in the North Atlantic and Indonesian region and a minimum near the Southern tip of India (Fig. 2.11). GNSS do not have direct access to geoid-related (mean sea level) heights but can only obtain the height with respect to a reference ellipsoid from the navigation solution. For conversion of ellipsoidal heights to mean sea level, a database of precomputed geoid undulations can be used within a GNSS receiver. As an example, [2.44] provides tabular geoid heights on a $10^\circ \times 10^\circ$ longitude/latitude grid. The geoid height at any user location can then be ob-

tained through interpolation using a weighted average of the nearest four grid points with a root-mean-square accuracy of better than 4 m. Higher accuracy would require a finer grid and a more accurate geoid model, such as EGM2008.

2.3.2 Global Reference System

The definition of a global terrestrial reference system (or, *terrestrial reference system*, TRS) began in earnest with the advent of Earth-orbiting satellites that enabled a realization of the center of mass and thus a natural origin for the coordinate system. Other names are conventional terrestrial reference system and geocentric terrestrial reference system. The roots of efforts to define a global system, however, can be traced back to the turn of the last century. In 1899, the *International Latitude Service (ILS)* was established by the *International Association of Geodesy (IAG)* to conduct astronomic latitude observations that monitor the motion of Earth's rotation axis relative to the Earth (polar motion, Sect. 2.5.3). By observing and disseminating this motion, latitudes and longitudes obtained by observing the stars could be corrected so that they refer to a fixed global terrestrial system.

In 1960, it was decided at the General Assembly of the *International Union of Geodesy and Geophysics (IUGG)* to adopt as terrestrial pole the average of the

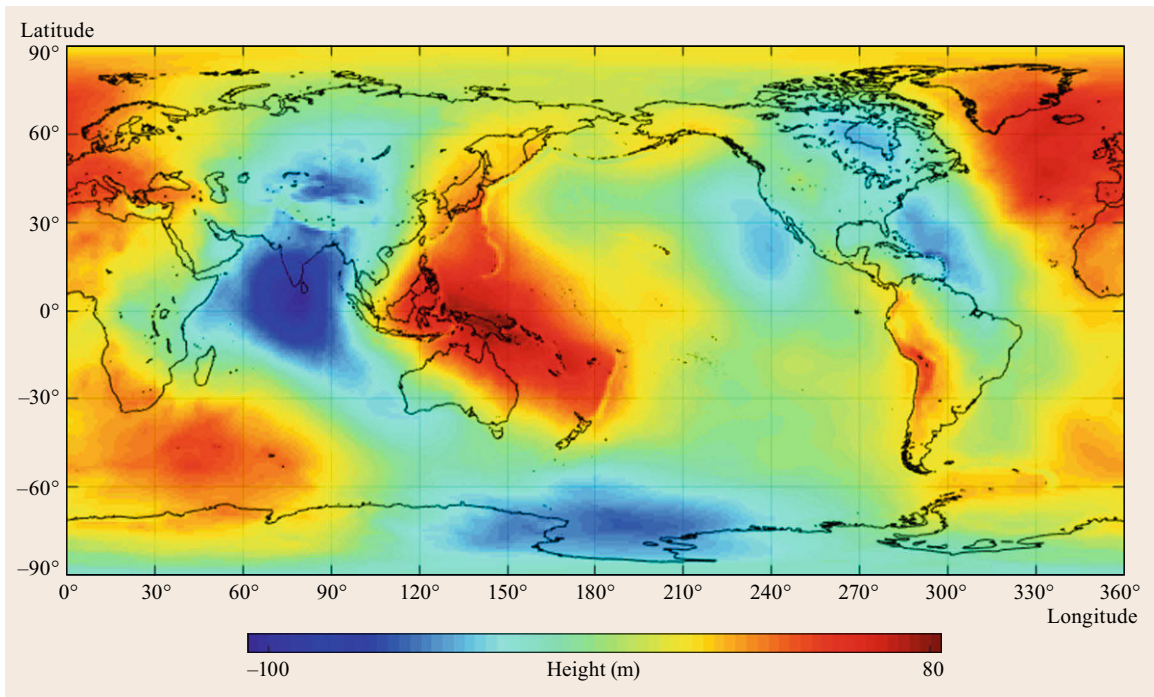


Fig. 2.11 Geoid heights relative to the Earth ellipsoid (courtesy of Th. Fecher)

positions of the true celestial pole on the Earth during the period 1900–1905 (a 6 year period over which the Chandler period of 1.2 years would repeat five times; Sect. 2.5.3). This average was named the *Conventional International Origin* (CIO).

The global reference meridian, or, origin for longitudes, originally defined astronomically as the meridian through the Greenwich observatory, near London, England, was realized by an average, as implied by the longitudes of many observatories around the world, corrected for polar motion and length-of-day variations.

These early conventions and procedures to define and realize a terrestrial reference system addressed astronomic *directions* only; no attempt was made to define a realizable origin, although implicitly it could be considered as geocentric. In 1984, the BIH, responsible until this time for monitoring the pole and the Greenwich meridian, defined the BIH Conventional Terrestrial System (CTS) (or also BTS) based on satellite laser ranging (SLR), VLBI, and other space techniques. With the inclusion of satellite observations, an (indirectly) accessible geocentric origin of the system could now be defined. New and better satellite and VLBI observations became available from year to year, and the BIH published new realizations of its system: BTS84, BTS85, BTS86, and BTS87. With the orientation of the TRS now defined by geometric satellite and space observations, the origin of geodetic longitudes, to be consistent with its astronomic counterpart (maintained for continuity in time keeping), is now about 102 m to the east of the Greenwich Observatory, which accounts for the local deflection of the vertical [2.45].

In 1988, the functions of monitoring the pole and the reference meridian were turned over to the newly established *International Earth Rotation Service* (IERS). The time service, originally also under the BIH, now resides with the BIPM. The new reference pole realized by the IERS, called the *International Reference Pole* (IRP), was adjusted to fit the BIH reference pole of 1967–1968 and presently is consistent with the CIO to within $\pm 0.03''$ (1 m).

The IERS, renamed in 2003 to *International Earth Rotation and Reference Systems Service* (retaining the same acronym), is responsible for defining and realizing both the *International Terrestrial Reference System* (ITRS) and the *International Celestial Reference System* (ICRS). In each case, an origin, an orientation, and a scale are defined among other conventions for the system. The system is then realized as a frame by the specification of these datum parameters and the coordinates of points worldwide. Since various observing systems (analysis centers and techniques) contribute to the overall realization of the reference system and since new realizations are obtained recurrently with improved

observation techniques and instrumentation, the transformations among various realizations of the system are of paramount importance. Especially, if one desires to combine data referring to realizations of different reference systems, or to different realizations of the same system, it is important to understand the coordinate relationships so that the data are combined ultimately in one consistent coordinate system.

The ITRS is defined by an orthogonal triad of right-handed, equally scaled axes with the following additional conventions:

1. The *origin* is geocentric, that is, at the center of mass of the Earth (including the mass of the oceans and atmosphere). Because measurement precision has reached the level of detecting variations in the center of mass due to terrestrial mass redistributions, the origin is defined as an average location of the center of mass and referred to an epoch.
2. The *scale* is defined by the SI meter, which is based on an adopted speed of light in vacuum and is connected to the definition of the SI second (Sect. 2.1).
3. The *orientation* is defined by the directions of the IRP and the reference meridian as given for 1984 by the BIH. Since it is now well established that Earth's crust (on which observing stations are located) is divided into tectonic plates that exhibit motion of the order of centimeters per year, it is further stipulated that the time evolution of the orientation of the reference system has no residual global rotation with respect to the crust (*no-net-rotation* condition). That is, even though the points on the crust, through which the system is realized, move with respect to each other, the net rotation of the system with respect to its initial definition should be zero.

The realization of the ITRS is the *International Terrestrial Reference Frame* (ITRF) and requires that three origin parameters, three orientation parameters, and a scale parameter must be identified with actual values. Each new ITRF of the system is named with the year that corresponds to the last available data incorporated in its realization. As of this writing (2015), the latest frame is ITRF2008 [2.46], and ITRF2013 is in preparation. The seven parameters are not observable without conventions (see below) and their specification is formulated by the IERS in terms of constraints imposed on the solution of coordinates from observations. Moreover, the constraints are cast in the form of a seven-parameter similarity transformation (commonly known as *Helmert* transformation) from an a priori given frame to the realized frame. The seven parameters include three translation parameters, T_i , that realize the origin; three

angle parameters, R_i , that realize the orientation; and, a scale change parameter, D , that realizes the scale

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{to}} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} + \begin{pmatrix} T_1 \\ T_2 \\ T_3 \end{pmatrix} + D \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} + \begin{pmatrix} 0 & -R_3 & +R_2 \\ +R_3 & 0 & -R_1 \\ -R_2 & +R_1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} \quad (2.29)$$

where the translation and rotation parameters are defined in Fig. 2.12. For example, if the origin of an existing frame is known to be the geocenter, then the next realization, based on new observations, can be related to the previous frame by constraining the translation to be zero. The transformation given by (2.29) is a linear approximation where, because of the small values of the transformation parameters, the neglect of second- and higher order terms affects coordinates at the subnanometer level.

Because these datum (transformation) parameters are determined for points on the Earth's crust (*crust-based frame*), and because the Earth as a whole is a dynamic entity, the parameters are associated with an epoch, t_0 , and, today, are supplemented with rates of

change, making the total number of parameters equal to 14. Thus, the i -th transformation parameter, β_i , is a function of time,

$$\beta_i(t) = \beta_{0,i} + \dot{\beta}_{0,i}(t - t_0), \quad (2.30)$$

and the 14 parameters are $\beta_{0,i}$ and $\dot{\beta}_{0,i}$ with $i = 1, \dots, 7$.

Whether the origin of a coordinate system is implied by a marker on the Earth's surface or accessed via distance measurements to Earth-orbiting satellites, it is defined by a convention, just like all other parts of the coordinate system. As such it is not, a priori, an observable quantity like a distance or an angle. This is the classic *datum defect* problem, well known in all types of surveying, where observations of distances and angles must ultimately be *related* to a point or direction that is fixed or defined by convention.

With satellite techniques, on the other hand, there is the advantage of knowing that the center of mass is the centroid for all orbits. Hence, the center of mass of the Earth serves as a natural origin point that, in theory, is accessible. That is, if the orbit is known, distance observations from points on the Earth's surface to points on the orbit are in a geocentric system, by definition. Due to observational error not all origin realizations, however, are identical as obtained by different analysis centers that, moreover, process different satellite data (such as satellite and lunar laser ranging [2.47, 48], GNSS [2.49], and Doppler data [2.50]). Generally, the most precise methods are based on SLR.

For the first ITRFs in the early 1990s, it was customary to relate all frames realized by particular analysis centers and/or satellite techniques to one of the SLR solutions from the Center for Space Research (CSR) in Austin, Texas, which was considered to be the best solution that accesses the center of mass and thus realizes the origin. The origins of solutions (i.e., realized coordinate systems) from other techniques, such as Doppler and GPS, were related by IERS to the ITRF origin through a translation determined by using stations that are common to both the CSR and the other solutions. For later ITRFs, a weighted average of selected SLR and GPS solutions was used to realize the origin. The origin of ITRF2000 was realized by a weighted average of the *most consistent SLR solutions* submitted to the IERS. With ITRF2005 and ITRF2008, the IERS used a time series over 13 years and 26 years, respectively, of reprocessed SLR data at selected, globally distributed sites to realize the origin.

Similarly, the scale was realized for the early ITRFs by the SLR solutions from the CSR analysis center, with the scale of other solutions transformed accordingly. For later realizations of scale, SLR was combined with VLBI, which accurately determines coordinate

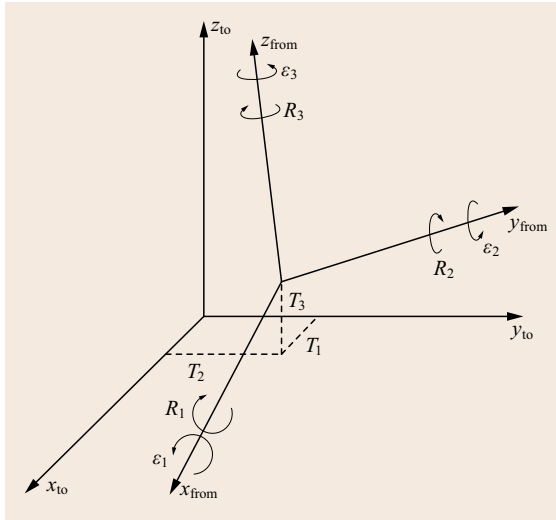


Fig. 2.12 Transformation parameters between coordinate frames. The similarity transformation (2.29) yields the coordinates of a point in a new frame \mathcal{R}_{to} that originates from the old frame $\mathcal{R}_{\text{from}}$ through translation of the origin by $-T_i$ ($i = 1, 2, 3$) and a *left-handed* rotation about the i -th axis by angle R_i . Rotation angles R_i comply with IERS conventions, whereas rotation angles $\epsilon_i = -R_i$ (corresponding to right-handed rotations), are used by the US National Geodetic Survey

differences of stations separated by large distances (several 1000 km) using observed directions to quasars [2.6, Chap. 4.2.2].

Satellite and space observational techniques contain no information on the absolute longitudinal orientation of a system. This orientation has no obvious natural reference and is completely arbitrary (the Greenwich meridian). One might argue that the orientation of the equatorial plane (or, equivalently, the polar direction), like the center of mass, is a natural reference that is accessible indirectly from astronomic observations, VLBI, and satellite tracking. However, the polar direction is complicated, a result of both polar motion with respect to the Earth's crust (Sect. 2.5.3), and precession and nutation with respect to the celestial sphere (Sect. 2.5.1). Besides this, the stations on the Earth's crust, which ultimately realize the ITRS, are in constant motion due to plate tectonics. Thus, the adopted convention for realizing the orientation of the ITRS is to ensure that each successive realization after 1984 is aligned with the orientation defined by the BIH in 1984 (with some early adjustments for different solutions of the Earth orientation parameters (Sect. 2.5.1)).

The methods of combining different solutions and introducing the constraints needed to address the datum defect (i. e., specifying origin, scale, and orientation) have become increasingly complicated as more data are assimilated and analysis centers employ different weighting schemes to account for the various observational accuracies. Relevant details may be found in the IERS Conventions of 2003 and 2010 and references therein, specifically also publications by [2.51, 52] and their references.

The model for the coordinates of any of the observing stations participating in the realization of ITRS is given by

$$\mathbf{x}(t) = \mathbf{x}_0 + (t - t_0)\mathbf{v}_0 + \sum_i \Delta\mathbf{x}_i(t), \quad (2.31)$$

where \mathbf{x}_0 and \mathbf{v}_0 are the vectors of the coordinates of the observing station and its velocity, defined for a particular epoch, t_0 . These are solved on the basis of observed coordinates, $\mathbf{x}(t)$, at time, t , using some type of observing system (e.g., SLR). The quantities, $\Delta\mathbf{x}_i$, are corrections applied by analysis centers to account for various, short wavelength, local geodynamic effects, such as solid Earth tides, ocean loading, and atmospheric loading (Sect. 2.3.5), with the objective of accounting for the nonconstant velocities. Details for the corresponding recommended models are provided by the IERS Conventions 2010 [2.6, Chap. 7]. The coordinate vector, \mathbf{x}_0 , and the linear velocity, \mathbf{v}_0 , for each participating station is provided by IERS as a result of

the assimilation of all data, and these represent the consequent realization of ITRS at epoch t_0 .

In the past, the linear velocity was modeled largely by the NNR-NUVEL1A tectonic plate motion model [2.32, 53, 54]. Thus,

$$\mathbf{v}_0 = \mathbf{v}_{\text{NUVEL1A}} + \delta\mathbf{v}_0, \quad (2.32)$$

where $\mathbf{v}_{\text{NUVEL1A}}$ is the velocity given as a set of rotation rates for the major tectonic plates, and $\delta\mathbf{v}_0$ is a residual velocity for the station. The major tectonic plates and site velocities predicted from a plate motion model are illustrated in Fig. 2.13. The newest ITRFs (since ITRF2000) appear to indicate significant departures of the station velocities \mathbf{v}_0 from the NNR-NUVEL1A model [2.55], which, however, does not impact the integrity of the ITRF.

2.3.3 Terrestrial Reference Systems for GNSS Users

The various navigation satellite systems have adopted specific reference systems for the provision of orbit information to their users. While the associated realizations may traditionally exhibit small offsets with respect to each other, GNSS providers are making continued effort to align the respective realizations with current versions of the ITRF.

In case of the US Global Positioning System, the World Geodetic System 1984 (WGS84, [2.56]) serves as the basis for orbit determination and broadcast ephemeris generation in the GPS control segment. WGS84 is the equivalent of the ITRS for the US Department of Defense (and includes also a global gravitational model). It is the evolution of previous reference systems, WGS60, WGS66, and WGS72 [2.57]. The corresponding reference frame for WGS84, as originally realized in 1987 on the basis mostly of satellite Doppler observations, agreed approximately with NAD83. The next realization, designated WGS84(G730), made use of observations from 12 GPS stations around the world and was aligned with the ITRF92 to an accuracy of about 20 cm in all coordinates. Here, G730 refers to GPS week 730 (Jan. 1994), the reference epoch of the WGS84 realization. Subsequent versions, known as WGS84(G873), WGS84(G1150) [2.58], and WGS84(G1674) [2.59], achieved continual improvements and are consistent, respectively, with ITRF94, ITRF2000, and ITRF2008 at the level of 10, 2, and 1 cm accuracy.

For the Russian Global'naja Nawigatsionnaja Sputnikowaya Sistema (GLONASS), the PZ-90 (Parametry Zemli – 90) system is employed. PZ-90 follows the same principles as the ITRS and WGS84, but is realized

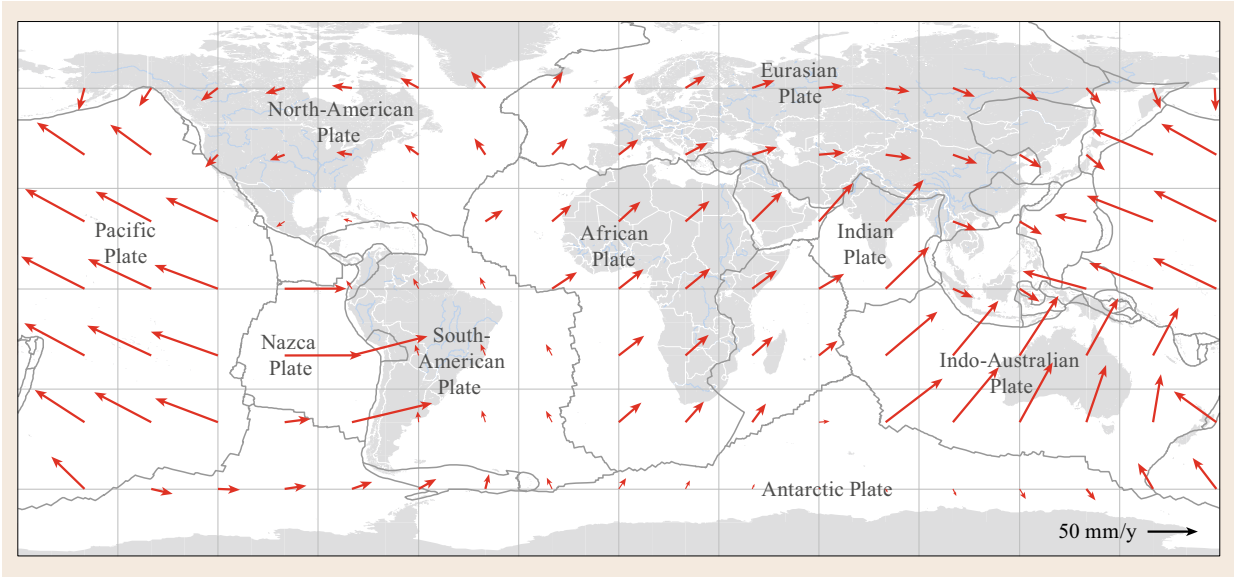


Fig. 2.13 Tectonic plates and predicted site velocities

through different reference stations and measurements. While the initial release of PZ-90 exhibited meter-level offsets from WGS84, the consistency was notably improved in 2007 with introduction of PZ-90.2 [2.60]. In early 2014, the GLONASS Control Segment finally switched to PZ-90.11 [2.61, 62], which offers a centimeter-level agreement with the latest ITRF versions.

Next to WGS84 and PZ-90, independent reference systems/frames are also employed for the BeiDou (China Geodetic Coordinate System 2000, CGS2000 [2.63]) as well as the European Galileo navigation system (Galileo Terrestrial Reference Frame, GTRF [2.64]).

2.3.4 Frame Transformations

The parameters of the Helmert similarity transformation (2.29) are determined in a weighted least-squares adjustment of the transformation model for the differences between coordinates of the same points in two frames. Table 2.2 lists the transformation parameters among the various IERS (and BIH) terrestrial reference frames since 1984. Due to the linear nature of the transformation, the reverse transformation is obtained by simply reversing the signs of the parameters. Also, the parameter values for a transformation between nonsuccessive frames are simply the accumulated values between the frames. However, especially for the later frames, the epoch of their validity must be considered. Rates of the parameters are given only since 1993 and (2.30) yields transformation parameters for

other than the listed epoch. Using the last row of Table 2.2 as an example, the translation in x between ITRF2005 and ITRF2008 at the epoch $t = 2000.0$ is given by

$$\begin{aligned} T_1(t) &= T_1(t_0) + \dot{T}_1 \cdot (t - t_0) \\ &= 0.05 \text{ cm} - 0.03 \text{ cm/y} \cdot (-5 \text{ y}) \\ &= 0.20 \text{ cm} . \end{aligned} \quad (2.33)$$

On the other hand, each of the determined parameters also has an associated uncertainty (given by the IERS, but not listed in Table 2.2, which should be properly included in any such calculation).

Table 2.3 lists transformation parameters from the original realization of WGS84 to ITRF90 as published by the IERS [2.65], as well as from recent ITRFs to NAD83(CORS96) as published by the National Geodetic Survey [2.67]. There is no origin, scale, and orientation change between NAD83(2011) and NAD83(CORS96). Again, uncertainties in the parameters are not listed. Also, the more recent realizations WGS84 are essentially equivalent to the correspondingly contemporary ITRF (Sect. 2.3.3).

Resolutions 1 and 4 of the 1991 IAG General Assembly [2.68] recommend that regional high-accuracy reference frames be tied to an ITRF, where such frames associated with large tectonic plates may be allowed to rotate with these plates as long as they coincide with an ITRF at some epoch. This procedure was adopted for NAD83, which for the conterminous United States and Canada lies (mostly) on the North American tectonic plate. This plate has global rotational motion estimated

Table 2.2 Transformation parameters among ITRF and BTS frames for use with the 7/14-parameter Helmert model (2.29) and (2.30). Time-dependent transformation parameters are provided from ITRF93 onward. Based on data from [2.6, 65, 66]

From	To	$T_1 \dot{T}_1$ (cm) (cm/y)	$T_2 \dot{T}_2$ (cm) (cm/y)	$T_3 \dot{T}_3$ (cm) (cm/y)	$R_1 \dot{R}_1$ (0.001'') (0.001''/y)	$R_2 \dot{R}_2$ (0.001'') (0.001''/y)	$R_3 \dot{R}_3$ (0.001'') (0.001''/y)	$D \dot{D}$ (10 ⁻⁸) (10 ⁻⁸ /y)	t_0
BTS84	BTS85	+5.4	+2.1	+4.2	-0.9	-2.5	-3.1	-0.5	1984
BTS85	BTS86	+3.1	-6.0	-5.0	-1.8	-1.8	-5.81	-1.7	1984
BTS86	BTS87	-3.8	+0.3	-1.3	-0.4	+2.5	+7.5	-0.2	1984
BTS87	ITRF0	+0.4	-0.1	+0.2	0.0	0.0	-0.2	-0.1	1984
ITRF0	ITRF88	+0.7	-0.3	-0.7	-0.3	-0.2	-0.1	+0.1	1988
ITRF88	ITRF89	+0.5	+3.6	+2.4	-0.1	0.0	0.0	-0.31	1988
ITRF89	ITRF90	-0.5	-2.4	+3.8	0.0	0.0	0.0	-0.3	1988
ITRF90	ITRF91	+0.2	+0.4	+1.6	0.0	0.0	0.0	-0.03	1988
ITRF91	ITRF92	-1.1	-1.4	+0.6	0.0	0.0	0.0	-0.14	1988
ITRF92	ITRF93	-0.2	-0.7	-0.7	-0.39	+0.80	-0.96	+0.12	1988
		-0.29	+0.04	+0.08	-0.11	-0.19	+0.05	0.0	
ITRF93	ITRF94	-0.6	+0.5	+1.5	+0.39	-0.80	+0.96	-0.04	1988
		0.29	-0.04	-0.08	+0.11	+0.19	-0.05	0.0	
ITRF94	ITRF96	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1997
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ITRF96	ITRF97	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1997
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ITRF2000	ITRF2005	-0.01	+0.08	+0.58	0.0	0.0	0.0	-0.040	2000
		+0.02	-0.01	+0.18	0.0	0.0	0.0	-0.008	
ITRF2005	ITRF2008	+0.05	+0.09	+0.47	0.0	0.0	0.0	-0.094	2005
		-0.03	0.00	0.00	0.0	0.0	0.0	0.0	

Table 2.3 Transformation parameters for other terrestrial reference frames for use with the 7/14-parameter Helmert model (2.29) and (2.30). Note that $\varepsilon_1 = -R_1$, $\varepsilon_2 = -R_2$, $\varepsilon_3 = -R_3$ (after [2.62, 65, 67])

From	To	$T_1 \dot{T}_1$ (cm) (cm/y)	$T_2 \dot{T}_2$ (cm) (cm/y)	$T_3 \dot{T}_3$ (cm) (cm/y)	$\varepsilon_1 \dot{\varepsilon}_1$ (0.001'') (0.001''/y)	$\varepsilon_2 \dot{\varepsilon}_2$ (0.001'') (0.001''/y)	$\varepsilon_3 \dot{\varepsilon}_3$ (0.001'') (0.001''/y)	$D \dot{D}$ (10 ⁻⁸) (10 ⁻⁸ /y)	t_0
WGS72	ITRF90	-6.0	+51.7	+472.3	+18.3	-0.3	-547.0	+23.1	1984
WGS84 ^a	ITRF90	-6.0	+51.7	+22.3	+18.3	-0.3	+7.0	+1.1	1984
PZ-90	PZ-90.02	-107	-3	+2	0	0	-130	-22	2002
PZ-90.02	WGS-84(1150)	-36	+8	+18	0	0	0	0	2002
PZ-90.11	ITRF2008	-0.3	-0.1	0.0	+0.019	-0.042	+0.002	0.0	2010
ITRF96	NAD83(CORS96)	+99.1	-190.7	-51.3	+25.8	+9.7	+11.7	0.0	1997
		0.0	0.0	0.0	+0.053	-0.742	-0.032	0.0	
ITRF97	NAD83(CORS96)	+98.9	-190.7	-50.3	+25.9	+9.4	+11.6	-0.09	1997
		+0.07	-0.01	+0.19	+0.067	-0.757	-0.031	-0.02	
ITRF2000	NAD83(CORS96)	+99.6	-190.1	-52.2	+25.9	+9.4	+11.6	+0.06	1997
		+0.07	-0.07	+0.05	+0.067	-0.757	-0.051	-0.02	

^a original realization; for more recent realizations, see text.

according to the NNR-NUVEL1A model [2.54] by the rates

$$\Omega_x = +0.000258 \cdot 10^{-6} \text{ rad/y} = +0.053 \text{ mas/y}$$

$$\Omega_y = -0.003599 \cdot 10^{-6} \text{ rad/y} = -0.742 \text{ mas/y}$$

$$\Omega_z = -0.000153 \cdot 10^{-6} \text{ rad/y} = -0.032 \text{ mas/y}$$

(2.34)

which explain the rotation parameter rates between NAD83 and ITRF in Table 2.3.

Coordinates of a control point in any particular frame are listed in terms of the Cartesian vector \mathbf{x}_0 and a velocity $\dot{\mathbf{x}}_0$, both at a given epoch t_0 so that at any other epoch the coordinates within that frame are

$$\mathbf{x}(t) = \mathbf{x}_0 + \mathbf{v}_0(t - t_0). \quad (2.35)$$

Transformation between frames and epochs requires consideration of both the point velocity within a frame and the velocity of the transformation parameters. Thus,

$$\mathbf{x}_{\text{from}}(t_0) \xrightarrow{\beta_0} \mathbf{x}_{\text{to}}(t_0) \xrightarrow{\dot{\mathbf{x}}_0} \mathbf{x}_{\text{to}}(t), \quad (2.36)$$

or, also,

$$\mathbf{x}_{\text{from}}(t_0) \xrightarrow{\dot{\mathbf{x}}_{\text{from}}(t_0)} \mathbf{x}_{\text{from}}(t) \xrightarrow{\beta(t)} \mathbf{x}_{\text{to}}(t). \quad (2.37)$$

Transformations (2.36) and (2.37) are equivalent if the point and frame velocities are related according to

$$\dot{\mathbf{x}}_{\text{to}} = \dot{\mathbf{x}}_{\text{from}} + \dot{\mathbf{T}} + \dot{\mathbf{D}}\mathbf{x}_{\text{from}} + \dot{\boldsymbol{\Omega}}\mathbf{x}_{\text{from}}, \quad (2.38)$$

where

$$\boldsymbol{\Omega} = \begin{pmatrix} 0 & -R_3 & +R_2 \\ +R_3 & 0 & -R_1 \\ -R_2 & +R_1 & 0 \end{pmatrix}, \quad (2.39)$$

which is the time derivative (neglecting second and higher order terms) of (2.29).

For most points within a regional frame, such as NAD83, the within-frame velocity of a point is small compared to the velocity of that same point in the ITRF, since, in this example, most of the velocity within the ITRF is due to the motion of the North American plate, and the NAD83 rides along with that plate. However, points on another plate within the NAD83 frame, such as points on the West Coast that are on the Pacific Plate, experience significant motion within the frame.

2.3.5 Earth Tides

Because the Earth is not a rigid body, the coordinates of points on its surface change in time in response to forces that deform its crust. The largest of these is due to the gravitational attractions of the Sun and Moon, which not only create the familiar periodic motion of the ocean tides, but also deform any point on (or below) the elastic Earth. The resulting periodic motion is called the *Earth tide* or *body tide*. Furthermore, because of the ocean tides, there is a secondary loading effect that deforms the crust especially at points near coastal areas. These tidal deformations are part of the corrections $\Delta\mathbf{x}_i(t)$ in (2.31).

In addition to the tide-induced corrections, there are other environmental effects, such as subsidence or uplift due to natural geophysical effects (earthquakes, post-glacial rebound) or anthropogenic activities (sub-surface mineral and water extraction), and due to local hydrological effects (seasonal, secular, and episodic

changes). These are site specific and dependent on local models.

The starting point for computing the tidal effect is the *tidal potential*, which accounts for the relative gravitational attraction of the Sun and Moon (other bodies have negligible effect). It is defined as the residual potential after removing the potential associated with the gravitational acceleration that is constant at all material points of the Earth. Assuming that the gravitational effect of a celestial body, B (e.g., the Sun, \odot , or Moon, ☾), may be approximated as that of a point mass at location, (r_B, ϕ_B, λ_B) , in the terrestrial reference system, the principal tidal potential at (r, ϕ, λ) and time t is given by [2.2, p. 132]

$$V^{(B)}(r, \phi, \lambda, t) = \frac{GM_B}{5r_B} \left(\frac{r}{r_B} \right)^2 \times \sum_{m=0}^2 \bar{P}_{2,m}(\sin \phi) \bar{P}_{2,m}(\sin \phi_B) \cos(mt_B), \quad (2.40)$$

where G is Newton's gravitational constant, M_B is the mass of the body, $\bar{P}_{2,m}$ is a second-degree, m -th order, fully normalized, associated Legendre function [2.42] and

$$t_B = t_{\Upsilon}^G + \lambda - \alpha_B \quad (2.41)$$

is the hour angle of the body, combining λ , the Greenwich sidereal time, t_{Υ}^G , and the right ascension, $\alpha_B = t_{\Upsilon}^G + \lambda_B$, of the body (Fig. 2.16). The coordinates, r_B, ϕ_B, α_B , and t_{Υ}^G are functions of time describing both the orbit of the body around the Earth and Earth rotation.

Equation (2.40) separates the long-period tides ($m = 0$) that have annual, semiannual, monthly, and fortnightly periods due to the orbital motion of the Earth and Moon, and the diurnal ($m = 1$) and semidiurnal ($m = 2$) tides due to Earth's rotation. In fact, (2.40) is an approximation that includes only the second-degree harmonics of the potential. Including third-degree harmonics, having the much smaller factor, $(r/r_B)^3$, and Legendre functions, $\bar{P}_{3,m}$, $m = 0, 1, 2, 3$, would introduce terdiurnal periods.

The tidal potential includes a permanent tide potential that is the average over time. Only the $m = 0$ term contributes and is calculated by averaging $\phi_B(t)$ over one orbit assuming r_B is constant [2.69],

$$V_{\text{perm}}^{(B)}(r, \phi) = \frac{3}{8} \frac{GM_B r^2}{r_B^2} (3 \sin^2 \phi - 1) \cdot \left(\sin^2 \varepsilon - \frac{2}{3} \right), \quad (2.42)$$

where ε is the obliquity of the ecliptic (Sect. 2.4). Since the potential is a scalar function, the law of superposition says that the tidal potential due to all bodies is the sum of the individual tidal potentials; thus, $V = V^{\oplus} + V^{\odot}$.

The tidal deformation of points on the Earth derives heuristically from the elasticity of the Earth and *Hooke's law*, which states that a displacement of the end of an elastic spring (the Earth's surface in this case) is linearly proportional to an applied force (the gravitational pull by the body). The gravitational pull (per unit mass) is the gradient of the potential; and, as a vector it creates a three-dimensional deformation in the radial and locally horizontal directions (Fig. 2.7),

$$\begin{pmatrix} \Delta u \\ \Delta v \\ \Delta w \end{pmatrix} = \begin{pmatrix} \ell_2 \frac{a}{g_0} \frac{\partial V}{r \cos \phi \partial \lambda} \\ \ell_2 \frac{a}{g_0} \frac{\partial V}{r \partial \phi} \\ \frac{h_2}{2} \frac{a}{g_0} \frac{\partial V}{\partial r} \end{pmatrix}, \quad (2.43)$$

where Earth's equatorial radius, a , and an average value of Earth's gravity, g_0 , are introduced so that the *spring constants*, h_2 , ℓ_2 , are dimensionless (the subscript refers to the second-degree model of the tidal potential). Indeed, h_2 was postulated by A.E.H. Love in 1909 and has become known as a *Love number*. The factor of 2 is included here since Love originally assumed simple proportionality to the tidal potential. In fact, for points on a spherical Earth, $\partial V / \partial r = 2V/a$; see also [2.70] for a definition in terms of vector spherical harmonics that is adopted by the IERS. Likewise, ℓ_2 is called a *Shida number*, although both are now generically called Love numbers.

The displacements given by (2.43) include a component due to the permanent tide, (2.42); but, such a displacement is time invariant and cannot be observed. Although the IAG in 1984 recommended that station coordinates *not* be corrected for the permanent tidal deformation, the continuing practice is to apply the full tidal effect, thus placing the coordinates in a *tide-free system*, rather than the recommended *mean-tide system*, which only has time-varying components removed [2.6, p. 108].

Nominal values for the Love numbers are [2.6]

$$h_2 = 0.61 \quad \text{and} \quad \ell_2 = 0.085, \quad (2.44)$$

which yield, with the corresponding astronomical constants for the Moon and Sun, a permanent deformation

at the equator of $\Delta w_{\oplus}^{(0)} = 5.5$ cm and $\Delta w_{\odot}^{(0)} = 2.5$ cm. The diurnal variations with respect to these mean values and the simple model above amount to less than 20 cm for the Moon and less than 10 cm for the Sun.

The Love numbers depend strongly on the density and elastic properties of the Earth, including its liquid core, and to a lesser extent on its ellipticity and changes in Earth orientation (nutation and polar motion). In particular, the resonance with the nearly diurnal free wobble (free core nutation, Sect. 2.5.3) is significant and illustrates that the Love numbers are also frequency dependent. The simple model has been extended with various levels of sophistication to account for the deformations observed with VLBI; see [2.70–72], and references therein, and [2.6] that summarizes the recommended formulas.

The secondary effect on station positions, due to ocean loading, depends on the ocean tide model and is computed using a convolution of ocean tide height with a Green's function [2.73]. The effect can be several percent of the body tide effect within continents and several 10% near the coast [2.30]. Another source of variable loading comes from the atmospheric tides resulting from the diurnal heating by the Sun. These mm-to-cm effects can be computed from corresponding atmospheric tidal models based on worldwide barometric data.

The centrifugal acceleration associated with Earth's rotation changes at a point with changes in the direction of the rotation axis with respect to the crust (and thus the terrestrial reference system). This implies a further deformation for an elastic Earth with the periods of polar motion (Sect. 2.5.3). It is called the *pole tide* although the source is not an external gravitational field. The centrifugal acceleration, $\mathbf{a}_c = \nabla V_c$, may be associated with a centrifugal potential, V_c , whose residual relative to $V_c^{(0)} = 0.5\omega_{\oplus}^2 (x^2 + y^2)$, is shown by [2.71] to be, in first-order approximation

$$\delta V_c = -\frac{\omega_{\oplus}^2}{2} r^2 \sin 2\phi (x_p \cos \lambda - y_p \sin \lambda), \quad (2.45)$$

where x_p, y_p are the coordinates of the pole in the TRS (Sect. 2.5.3), measured in radians. This has the same form as the second-degree tidal potential (2.42) due to an extraterrestrial body; and, the corresponding crustal deformation is given by (2.43). Since $|x_p|, |y_p| \approx 0.2''$ relative to the current mean position, the vertical variation is of the order of 0.6 cm. The effect of ocean loading due to the pole tide, again, is site and ocean-basin model dependent and at the level of a millimeter [2.6, Chap. 7].

2.4 Celestial Reference System

Throughout history and today the ultimate reference system for positioning and navigation on and near the Earth, as well as within our solar system is an astrometric system. Its modern manifestation is the *celestial reference system* (CRS). By definition, it is an *inertial* system, which means that it is in free fall in the gravitational field of the universe and does not rotate. It is a system in which the laws of physics in the context of the theory of general relativity hold without requiring corrections for rotations. For geodetic purposes the CRS serves as the primal reference for positioning since it has no dynamics. Conversely, it is the system with respect to which we study the dynamics of the Earth as a rotating body. And, finally, it serves, of course, also as a reference system for astrometry.

The *celestial reference frame* (CRF) is the realization of the CRS based on a set of coordinates of objects on the celestial sphere. For this purpose the origin of the celestial sphere, and thus the CRS, is defined to be the *barycenter* of the solar system, which is the center of mass, as realized by the orbits of the planets. When appropriate or necessary, one also makes the distinction between the CRS and the *geocentric celestial reference system* (GCRS).

The origins, or zero points, of the celestial coordinates, declination and right ascension, have changed definition with a fundamental redefinition of the CRS in 1998. Prior to this time, the definition of the celestial reference system was tied directly to the dynamics of the Earth, whereas, today it is defined almost completely independent of the Earth, although the difference in realizations is defined to be minimal for the sake of continuity. The traditional system refers to two natural directions, the mean direction of Earth's spin axis, or the *north celestial pole* (NCP), and the direction of the *north ecliptic pole* (NEP), which is perpendicular to the mean ecliptic plane defined by Earth's orbit around the Sun (Fig. 2.14).

A point where the ecliptic crosses the celestial equator on the celestial sphere is called an equinox. At the

vernal equinox, Υ , the Sun crosses the celestial equator from south to north as viewed from the Earth. It is the point on the Earth's orbit when Spring starts in the Northern Hemisphere. The angle between the celestial equator and the ecliptic is the *obliquity of the ecliptic*, approximately $\varepsilon = 23.44^\circ$.

The direction of the vernal equinox defines the traditional origin for right ascension and the celestial equator is the origin for declination, as shown in Fig. 2.15. The system of celestial coordinates is also known as the *equatorial right ascension system*. The first and third axes of this system are, respectively, the directions of the vernal equinox and the NCP, which are perpendicular since the vernal equinox lies in the equatorial plane. The second axis is perpendicular to the other two axes so as to form a right-handed system. The intersection of the celestial sphere with the plane that contains both the third axis and a celestial object is called the *hour circle* of the object (Fig. 2.16). The right ascension system is the basis for the celestial reference system, where one must further fix the axes since both the vernal equinox and the NCP are dynamic directions that vary in time due to gravitational effects on Earth's rotational direction and its orbit.

The relationship between the right ascension and longitude on the Earth is illustrated in Fig. 2.16 under the assumption that the terrestrial pole and the NCP have the same direction (Sect. 2.5.1). The name, hour circle, refers to the convention that the right ascension of an object is also given in terms of a sidereal time interval (Sect. 2.1.3), where 15° of right ascension is equivalent to 1 h of sidereal time.

In order to define a reference system, it was necessary to establish the theory of motion of the NCP and the equinox, called the theory of nutation and precession (Sect. 2.5). Moreover, the realization of the reference system was stamped with an epoch for which it was valid; it was typical to determine a new realization every 25 years to reflect the dynamics of the reference system, as well as any updates in the theories

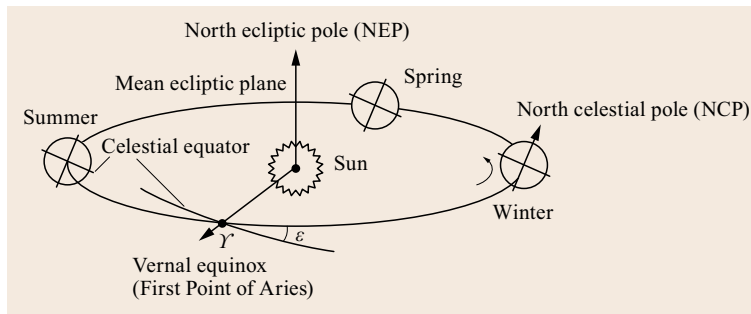


Fig. 2.14 Mean ecliptic plane (seasons are for the Northern Hemisphere). and natural directions for the celestial reference system

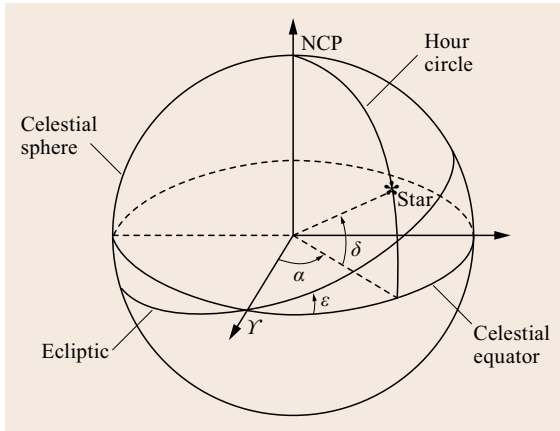


Fig. 2.15 Celestial coordinates, α , δ , in the equatorial right ascension system

of motion [2.5, p. 167]. The last such realization was the **FK5** (Fundamental Catalog No. 5) of stars referring to the best computed equinox and NCP for the epoch, J2000. The origins of right ascension and declination were determined indirectly from an adjustment of observed coordinates of celestial objects and their proper motions (in other words, the equinox is not observed directly).

The change in definition of the CRS adopted by the International Astronomical Union (IAU) in the 1990s was enabled by the then established history of very accurate observations of quasars (quasi-stellar radio sources) using the technique of Very Long Baseline Interferometry (VLBI, [2.39, Chap. 11.1]). Since these beacons are at such great distances that no proper motion can be detected, that is, they have no perceptible rotation on the celestial sphere, they may be used to define an inertial system.

This new definition of the CRS represents a change as fundamental as that which transferred the origin of the regional terrestrial reference system (i. e., the horizontal geodetic datum) from a monument on Earth's surface to the geocenter. By relying strictly on geometrically defined points on the celestial sphere, the definition of the CRS has changed from a *dynamic* system to a *kinematic* (or geometrical) system. The axes of the celestial reference system are still (close to) the NCP and vernal equinox, but are not defined dynamically in connection with Earth's motion. Rather they are tied to the defining set of quasars whose coordinates are given with respect to these axes. Moreover, there is no need to define an epoch of reference, because (presumably) these directions will never change in inertial space (at least in the foreseeable future of mankind).

The IERS officially created the *International Celestial Reference System* (ICRS) starting in 1998 based

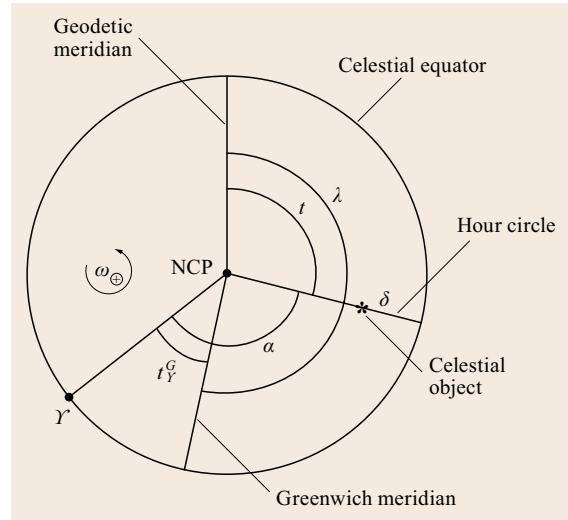


Fig. 2.16 Relationship between right ascension and longitude (idealized without polar motion). The meridian of a terrestrial point and the Greenwich meridian rotate relative to the vernal equinox due to Earth's rotation rate, ω_{\oplus} . The hour angle, t_Y^G , is also the Greenwich sidereal time

on 212 defining sources, which then also constitute the realization of the system, the *International Celestial Reference Frame* (ICRF). An additional 396 *candidate* or other less well observed sources were used as additional ties to the reference frame. The origin of the ICRS is defined to be the center of mass of the solar system (*barycentric* system) and is realized by observations of planets and other bodies in the solar system (such as the Jet Propulsion Laboratory (JPL) development ephemerides) in the framework of the theory of general relativity. These dynamical planetary ephemerides are aligned with the ICRF to high accuracy [2.6, Chap. 3].

By Recommendation VII of the 1991 IAU General Assembly, the NCP and equinox of the ICRS are supposed to be close to the mean dynamical pole and equinox of J2000.0. Furthermore, the adopted pole and equinox for ICRS should be consistent with the directions realized for FK5. Specifically, the origin of right ascension for FK5 was originally defined on the basis of the mean right ascension of 23 radio sources from various catalogs, with the right ascension of one particular source fixed to its FK4 value, transformed to J2000.0. Similarly, the FK5 pole was based on its J2000.0 direction defined using the 1976 precession and 1980 nutation series (Sect. 2.5). The FK5 directions are estimated to be accurate to ± 50 mas (milli-arcsec) for the pole and ± 80 mas for the equinox; and, it is now known, from improved observations and dynamical models, that the ICRS pole

and equinox are close to the mean dynamical equinox and pole of J2000.0, well within these tolerances. The precise transformation to a dynamical system is a *frame bias* that is included in the modern formulations of the transformations between the celestial and terrestrial reference frames (Sect. 2.5). This bias is well determined and of the order of 10 mas [2.6, 74].

As the VLBI measurements of the quasars improve, the orientation of the ICRF will be adjusted with the constraint that it has no net rotation with respect to previous realizations (analogous to the ITRF). The original realization is designated ICRF1; and, it was extended in 1999 and again in 2002 with additional objects ob-

served by VLBI, totaling 667 and 717, respectively. The next significant realization, designated ICRF2, was constructed in 2009, where now 295 quasars define the system (being more stable and better distributed in the sky than for ICRF1), and which also includes 3119 secondary extragalactic sources.

Aside from VLBI, the principal realization of the ICRS is through the Hipparcos catalog, based on recent observations of some 120 000 well-defined stars using the Hipparcos (High Precision Parallax Collecting Satellite), optical, orbiting telescope. This catalog is tied to the ICRF with an accuracy of about 0.6 mas in each axis. Additional catalogs for up to 100 million stars are described by [2.6].

2.5 Transformations Between ICRF and ITRF

The transformation from the CRF to the terrestrial reference frame requires an understanding of the dynamics of Earth rotation and its orbital motion, as well as the effects of observing celestial objects on a moving and rotating body such as the Earth. Even though the new definition of the celestial reference system (Sect. 2.4) no longer relies on models for the dynamics of the Earth's pole and the vernal equinox, but because the terrestrial system is fixed to the Earth, any transformation between celestial and terrestrial frames still does depend explicitly on these dynamics.

The description of the transformation, comprising *Earth orientation parameters*, has also changed with the adoption of the new system definition. Here, both the traditional description and the modern transformation are treated, where the traditional one is perhaps a bit more accessible in terms of physical intuition, whereas, the latter tends to hide these concepts. Furthermore, the new approach implements certain nuances necessary for an unambiguous definition of Earth rotation. Thus, the following starts with the traditional approach and evolves into the modern transformation formulas.

The theoretical description of Earth's *dynamics* in inertial space requires a system of time, and dynamic time, being theoretically the most uniform in scale (Sect. 2.1) is the natural choice. Because many of the dynamics vary on scales of years or longer, the time variable, τ , is expressed typically as a (unit-less) fraction of a Julian century relative to a fixed epoch

$$\tau = \frac{t - t_0}{36\,525}, \quad (2.46)$$

where $t_0 = 2451545.0$ is the Julian day number for J2000.0 and t is the Julian day number of the epoch of date.

2.5.1 Orientation of the Earth in Space

The gravitational interaction of the Earth with the other bodies of the solar system, including primarily the Moon and the Sun, but also the planets, cause Earth's orbital motion to deviate from the simple Keplerian motion of two point masses in space. Also, because the Earth is not a perfect homogeneous sphere, its rotation is affected likewise by the gravitational action of the bodies in the solar system. If there were no other planets (only the Earth/Moon system) then the orbit of the Earth/Moon system around the Sun would be essentially a plane fixed in space. This plane defines the ecliptic (Sect. 2.4). But the gravitational forces of the planets cause this ecliptic plane to behave in a dynamic way, called *planetary precession*.

If the obliquity of the ecliptic were zero (or the Earth were not flattened at its poles), then there would be no gravitational torques due to the Sun, Moon, and planets acting on the Earth's bulging equator. But since $\varepsilon \neq 0$ and $f \neq 0$, these celestial bodies (primarily, the Sun and Moon) do cause a precession of the equator and, hence, the pole, that is known as *luni-solar precession* and *nutation*, depending on the period of the motion [2.75]. That is, the equatorial bulge of the Earth and its tilt with respect to the ecliptic allow the Earth to be torqued by the gravitational forces of the Sun, Moon, and planets, since they all lie approximately on the ecliptic plane. Planetary precession together with luni-solar precession is known as *general precession*.

The complex dynamics of the precession and nutation is a superposition of many periodic motions originating from the myriad of periods associated with the orbital dynamics of the corresponding bodies. Conventionally, the period of 18.6 years associated with

the longest lunar cycle separates nutation from precession, where the latter can be described virtually as a secular motion of the pole and equinox owing to their fundamental respective periods of about 25 800 and 28 100 years. The periods of nutation depend primarily on the orbital motion of the Moon relative to the orbital motion of the Earth. The most recent models for nutation also contain short-periodic effects due to the relative motions of the planets. In terms of transformations, precession has been viewed as an accumulation of *mean* motion over a time interval, whereas, nutation is thought of as the correction, or residual, that transforms from the mean to the true location of the pole and equinox at a particular instant in time.

The theory for determining the motions of the coordinate reference directions was developed by Simon Newcomb at the turn of the twentieth century. Its basis lies in celestial mechanics and involves the n -body problem for planetary motion, for which no analytical solution exists. Instead, iterative, numerical procedures have been developed and formulated [2.76].

Precession

Planetary precession may be described by two angles, π_A and Π_A , where the subscript, A , refers to the *accumulated* angle from some fixed epoch, t_0 , to some other epoch, t . Figure 2.17 shows the geometry of the motion of the ecliptic due to planetary precession from t_0 to t . The pictured ecliptics and equator are fictitious in the sense that they are affected only by precession, but not nutation, and as such are called *mean ecliptic* and *mean equator*. The angle, π_A , is the angle between the mean ecliptics at t_0 to t ; while, Π_A is the ecliptic longitude of the point, M , on the celestial sphere, which identifies the axis of rotation of the ecliptic due to planetary precession. The vernal equinox at t_0 is denoted by Υ_0 . Expressions for π_A and Π_A are truncated time series based on the celestial dynamics of the planets.

The *luni-solar precession*, on the other hand, also depends on the geophysical parameters of the Earth. Due to the more complicated nature of the Earth's shape and internal constitution, no analytic formula based on theory has been used for this part of the precession. Instead, Newcomb gave an empirical parameter, (now)

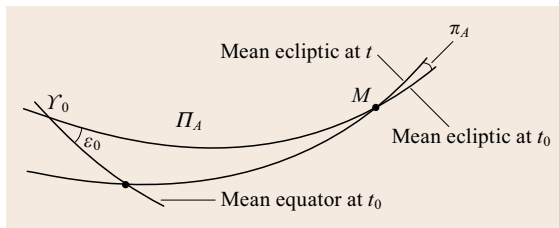


Fig. 2.17 Planetary precession

called *Newcomb's precessional constant*, based on observed rates of precession. In fact, this *constant* rate is not strictly constant, as it depends slightly on time through a general relativistic term called the *geodesic precession* [2.77]. Newcomb's precessional constant depends on Earth's moments of inertia and enters in the dynamical equations of motion for the celestial equator due to the gravitational torques of the Sun and Moon.

Figure 2.18 shows the accumulated angles of planetary and luni-solar precession near the vernal equinox. The precession angles, ψ_A and χ_A , respectively, describe the motion of the mean vernal equinox along the mean ecliptic (luni-solar precession) and along the mean equator (planetary precession).

Due to their virtually secular variation over the near term (few thousands of years), the planetary and luni-solar precessional angles are expressed as polynomials in time, formulated with a certain set of adopted constants and a dynamical theory. The developments of Newcomb and his contemporaries was reformulated and extended in precision by [2.77] based on constants adopted by the IAU. This was further extended in precision and updated in 2000 and again in 2006 by the IAU based on the works of [2.58, 78]. The resulting model, designated the IAU 2006 precession, includes, for example, expressions

$$\begin{aligned}\psi_A &= 5038.481507''\tau - 1.0790069''\tau^2 \\ &\quad - 0.00114045''\tau^3 + \dots \\ \chi_A &= 10.556403''\tau - 2.3814292''\tau^2 \\ &\quad - 0.00121197''\tau^3 + \dots,\end{aligned}\quad (2.47)$$

for the angles ψ_A and χ_A , where τ is given by (2.46) and where fourth- and fifth-order terms are omitted here for brevity. The linear parts then give the instantaneous rates of precession at t_0 . The rate of luni-solar precession of the vernal equinox along the mean ecliptic is

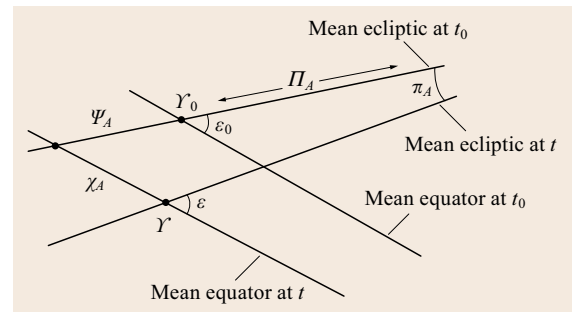


Fig. 2.18 General precession of the vernal equinox. Planetary precession along the mean equator is indicated by the angle, χ_A , and ψ_A denotes the luni-solar precession along the mean ecliptic (not to scale)

as

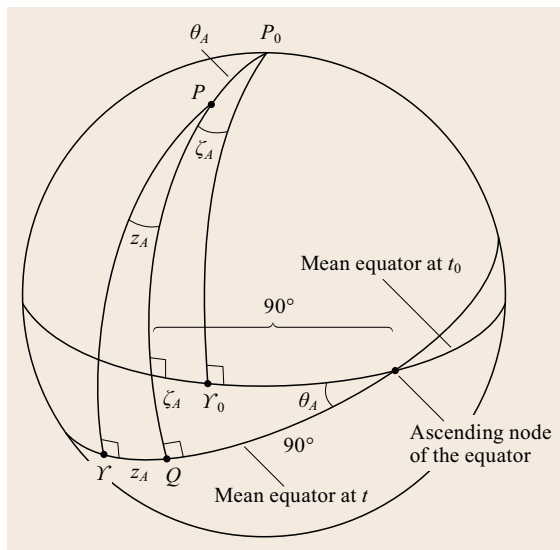
$$\mathbf{r}_0 = \begin{pmatrix} \cos \alpha_0 \cos \delta_0 \\ \sin \alpha_0 \cos \delta_0 \\ \sin \delta_0 \end{pmatrix} \quad (2.50)$$

and \mathbf{r}_m , analogously. Then, the transformation between the two frames is achieved by the rotations

$$\begin{aligned}\mathbf{r}_m &= \mathbf{R}_3(-z_A)\mathbf{R}_2(+\theta_A)\mathbf{R}_3(-\zeta_A)\mathbf{r}_0 \\ &= \mathbf{P}\mathbf{r}_0\end{aligned}\quad (2.51)$$

where it is noted that the great circle arc, $\widehat{P_0PQ}$, intersects both mean equators of t_0 and of t at right angles because it is an hour circle with respect to both poles, P_0 and P . $\mathbf{R}_j(\alpha_j)$ denotes the usual orthogonal rotation matrix for a rotation by the angle, α_j , about the j -th axis of a right-handed Cartesian coordinate system (Table 2.4). \mathbf{P} is called the *precession transformation matrix*.

The precessional elements for the IAU 2006 model [2.78, 81] are given by

$$\begin{aligned}\xi_A &= 2.650545'' + 2306.083227''\tau \\ &\quad + 0.2988499''\tau^2 + 0.01801828''\tau^3 \\ &\quad - 0.5971'' \cdot 10^{-6}\tau^4 - 3.173'' \cdot 10^{-7}\tau^5 \\ z_A &= -2.6505453'' + 2306.0771813''\tau \\ &\quad + 1.09273483''\tau^2 + 0.018268373''\tau^3 \\ &\quad - 28.596'' \cdot 10^{-6}\tau^4 - 2.904'' \cdot 10^{-7}\tau^5 \\ \theta_A &= 2004.191903''\tau - 0.4294934''\tau^2 \\ &\quad - 0.041822''\tau^3 - 7.089'' \cdot 10^{-6}\tau^4 \\ &\quad - 1.274'' \cdot 10^{-7}\tau^5, \end{aligned} \tag{2.52}$$


where, however, these expressions do not include the frame bias introduced with the change in celestial reference system definition (Sect. 2.4).

Nutation

Since the nutations are primarily due to the luni-solar attractions, they are modeled firstly in terms of the ecliptic coordinates of the Sun and Moon. Traditionally, the nutations are expressed by two angles, $\Delta\epsilon$ and $\Delta\psi$, that, respectively, describe the change (from mean to

Table 2.4 Elementary rotation matrices. Multiplication of a coordinate vector referred to a frame $\mathcal{R}_{\text{from}}$ by matrix \mathbf{R}_i provides the coordinates of the same vector in a frame \mathcal{R}_{to} , which is obtained from $\mathcal{R}_{\text{from}}$ by a right-handed rotation by angle α about the i -th axis (after [2.79, 80])

Rotation about x-axis:	Rotation about y-axis:	Rotation about z-axis:
$\mathbf{R}_1(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & +\cos \alpha & +\sin \alpha \\ 0 & -\sin \alpha & +\cos \alpha \end{pmatrix}$	$\mathbf{R}_2(\alpha) = \begin{pmatrix} +\cos \alpha & 0 & -\sin \alpha \\ 0 & 1 & 0 \\ +\sin \alpha & 0 & +\cos \alpha \end{pmatrix}$	$\mathbf{R}_3(\alpha) = \begin{pmatrix} +\cos \alpha & +\sin \alpha & 0 \\ -\sin \alpha & +\cos \alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}$

true) in the tilt of the equator with respect to the mean ecliptic, and the change (again, from mean to true) of the equinox along the mean ecliptic (Fig. 2.20). There is no need to transform from the mean to the true ecliptic since only the dynamics of the true equator are of interest. The true vernal equinox, Υ_T , is always defined to be on the mean ecliptic.

The *nutation in longitude*, $\Delta\psi$, is primarily caused by the ellipticities of the Earth's and Moon's orbits. The *nutation in obliquity*, $\Delta\epsilon$, is mainly due to the Moon's orbital plane being out of the ecliptic (by about 5.145°). Models for the nutation angles are given in the form

$$\begin{aligned}\Delta\psi &= \sum_{i=1}^n (a_i \sin A_i + a'_i \cos A_i) \\ \Delta\epsilon &= \sum_{i=1}^n (b_i \cos A_i + b'_i \sin A_i),\end{aligned}\quad (2.53)$$

where each amplitude, a_i, a'_i, b_i, b'_i , is a linear function of τ and the angle

$$A_i = n_{\ell,i}\ell + n_{\ell',i}\ell' + n_{F,i}F + n_{D,i}D + n_{\Omega,i}\Omega, \quad (2.54)$$

represents a linear combination of fundamental arguments (*Delaunay variables*, [2.82]) of the solar and lunar orbits:

- ℓ Mean anomaly of the Moon,
- ℓ' Mean anomaly of the Sun,
- F Mean longitude of the Moon minus the mean longitude of the Moon's ascending node,
- D Mean elongation of the Moon from the Sun,
- Ω Mean longitude of the ascending node of the Moon.

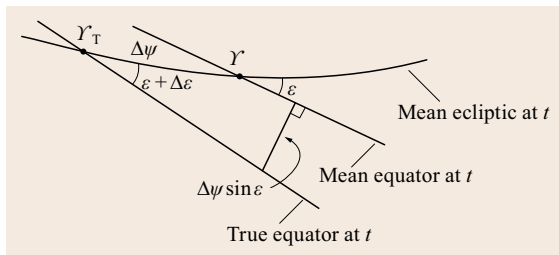


Fig. 2.20 Nutation elements, $\Delta\epsilon$ and $\Delta\psi$

The corresponding arguments are introduced for the planetary orbits in an expanded theory. The integer multipliers, $n_{\ell,i}, \dots, n_{\Omega,i}$, specify how these variables are combined in the argument, A_i .

The theory and series developed by [2.76] included $n = 69$ terms for $\Delta\psi$ and $n = 40$ terms for $\Delta\epsilon$. The subsequent theory and series [2.83] adopted by the IAU in 1980, which included modifications for a nonrigid Earth model [2.71] had $n = 106$ terms. The IAU1980 nutation model was replaced in 2003 by the new nutation model of [2.58], designated IAU2000A (2000B is an abbreviated, less precise version). This model accounts for the mantle anelasticity, the effects of ocean tides, electromagnetic couplings between the mantle, the fluid outer core, and the solid inner core, as well as various nonlinear terms not previously considered.

A slight revision of the model due to the new IAU 2006 precession model is designated the IAU2000A_{R06} nutation model, which has 1320 terms for $\Delta\psi$ and 1037 terms for $\Delta\epsilon$ ([2.6] and [2.84, Tables 5.3a,b]). Table 2.5 summarizes the largest of the nutation amplitudes and associated variables and parameters according to this model. The corresponding expressions for the Delaunay variables as low-order polynomials in τ are also given in [2.6, p. 67]. The periods of the nutations may be computed from the linear coefficients of the resulting polynomial expressions for the angle, A_i . The high-index angles, A_i , also include the longitudes of the planets. The frame bias (Sect. 2.4) is already incorporated in Table 2.5.

Figure 2.21 depicts the motion of the pole due to the dominant luni-solar precession combined with the largest of the nutation terms. This diagram also defines the *nutation ellipse* that describes the extent of the true motion with respect to the mean motion. The semi-axis of this ellipse that is orthogonal to the mean motion is the principal term in the nutation in obliquity and is also known as the *constant of nutation*. The values for it and for the other semi-axis, given by $\Delta\psi \sin \epsilon$ (Fig. 2.20), can be inferred from Table 2.5. The total motion of the pole on the celestial sphere is due to the superposition of the general precession and all the nutations.

The transformation at the epoch of date, t , from the mean frame to the true frame, referring to Fig. 2.20, is

Table 2.5 The dominant terms of the IAU2000A_{R06} series for nutation in longitude and obliquity, referred to the mean ecliptic of date. τ , as defined in (2.46) denotes the number of Julian centuries since 1.5 Jan 2000. Note that the index i does not correspond to the order of the IAU $\Delta\epsilon$ components.

i	Period (d)	$a_i (10^{-6}'')$		$b_i (10^{-6}'')$		$n_{\ell,i}$	$n_{\psi,i}$	$n_{F,i}$	$n_{D,i}$	$n_{Q,i}$
1	6798.4	-17 206 424.18	-17 418.82 τ	+9 205 233.10	+883.03 τ	0	0	0	0	+1
2	182.6	-1 317 091.22	-1369.60 τ	+573 033.60	-458.70 τ	0	0	+2	-2	+2
3	13.7	-227 641.81	+279.60 τ	+97 846.10	+137.40 τ	0	0	+2	0	+2
4	3399.2	+207 455.40	-69.80 τ	-89 749.20	-29.10 τ	0	0	0	0	+2
5	365.3	+147 587.70	+1181.70 τ	+7387.10	-192.40 τ	0	+1	0	0	0
6	27.6	+71 115.90	-87.20 τ	-675.00	+35.80 τ	+1	0	0	0	0
7	121.7	-51 682.10	-52.40 τ	+22 438.60	-17.40 τ	0	+1	+2	-2	+2
8	13.6	-38 730.20	+38.00 τ	+20 073.00	+31.80 τ	0	0	+2	0	+1
9	9.1	-30 146.40	+81.60 τ	+12 902.60	+36.70 τ	+1	0	+2	0	+2

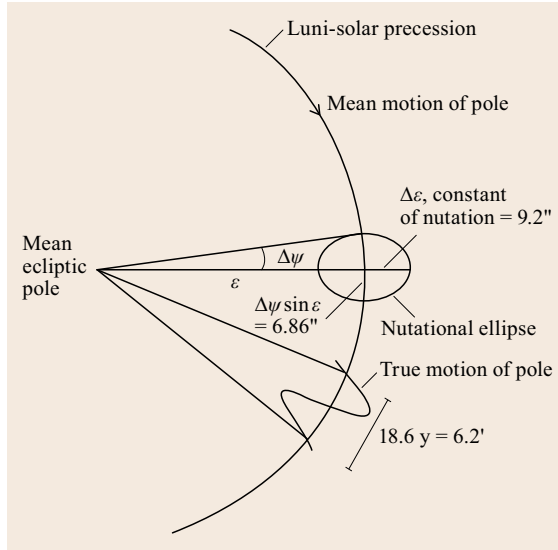


Fig. 2.21 Dominant components of the combined general precession and nutation of the pole on the celestial sphere

accomplished with the following rotations,

$$\begin{aligned} \mathbf{r} &= \mathbf{R}_1(-\epsilon - \Delta\epsilon)\mathbf{R}_3(-\Delta\psi)\mathbf{R}_1(\epsilon)\mathbf{r}_m \\ &= \mathbf{N}\mathbf{r}_m \end{aligned} \quad (2.55)$$

where ϵ is the mean obliquity at epoch, t , and the true right ascension and declination are related to \mathbf{r} as in (2.50).

The combined transformation due to precession and nutation from the epoch, t_0 , to the current epoch, t , is given by the combination of equations (2.51) and (2.55),

$$\mathbf{r} = \mathbf{N}\mathbf{P}\mathbf{r}_0. \quad (2.56)$$

The IAU 2006/2000A precession–nutation model is accurate to about 0.3 mas. For those seeking the highest accuracy and temporal resolution, small corrections

(called *celestial pole offsets*) obtained from continuing VLBI observations, may be applied to the nutation series. For example, the most recent model does not contain the diurnal motion due to the free-core nutation (FCN) caused by the interaction of the mantle and the rotating fluid outer core ([2.75]; see also Sect. 2.5.3). The IERS publishes differential elements in longitude, $\delta\psi$, and obliquity, $\delta\epsilon$, that can be added to the elements implied by the nutation series

$$\begin{aligned} \Delta\psi &= \Delta\psi(\text{model}) + \delta\psi \\ \Delta\epsilon &= \Delta\epsilon(\text{model}) + \delta\epsilon. \end{aligned} \quad (2.57)$$

2.5.2 New Conventions

The new definition of the celestial reference system (CRS, Sect. 2.4) was prompted not only by the ability to realize the system geometrically with accurate VLBI observations, but also by a critical analysis of the system conventions for the origin of right ascension [2.85, 86]. Specifically, by avoiding a dynamical definition of the CRS axes, there is no particular reason to use the vernal equinox on the mean ecliptic as an origin of right ascensions, especially because even in the mean it is a dynamical point on the celestial sphere. That is, as an origin point it rotates about the NCP due to the precessional rotation of the ecliptic. This must then be corrected when considering the rotation of the Earth with respect to inertial space (Greenwich sidereal time, or the hour angle at Greenwich of the vernal equinox; Sect. 2.1.3).

In 2000, the IAU adopted a set of resolutions that precisely adhered to a new, more accurate, and simplified way of dealing with the transformation between the celestial and terrestrial reference systems. The IERS, in 2003, similarly adopted the new methods based on these resolutions [2.87]. These were reinforced with IAU resolutions in 2006 and adopted as part of the IERS Conventions 2010. The true NCP, previously also called the celestial ephemeris pole (CEP) with a resolution of

the IAU in 1979, now is called the celestial intermediate pole (CIP), thus identifying it as a transition between celestial and terrestrial reference frames. The new conventions also revised the origin for right ascension in this intermediate frame so as to eliminate residual rotations not associated with Earth rotation, while also ensuring continuity with the previously defined origin. These profoundly new definitions solidify the paradigm of *kinematics* (rather than dynamics) upon which the celestial reference system is based. In addition, the description of precession and nutation is now combined in a single transformation from t_0 to t .

Suppose that the instantaneous pole, P , on the celestial sphere coincides with the reference pole, P_0 , at some fundamental epoch, t_0 . At the epoch of date, t , the position of P then has celestial coordinates as shown in Fig. 2.22. These coordinates are the co-declination, d , and the right ascension, E , with respect to the reference origin, Σ_0 . The true or instantaneous equator (the plane perpendicular to the axis through P) at time, t , intersects the reference equator (associated with P_0) at two nodes that are 180° apart. The hour circle of the node, N , is orthogonal to the great circle arc $\widehat{P_0P}$. Therefore, the right ascension of the ascending node of the equator is 90° plus the right ascension of the instantaneous pole, P . The origin for right ascension at the epoch of date, t , is defined kinematically under the condition that there is no rotation rate of the *instantaneous coordinate frame* about the pole due to precession and nutation. This is the concept of the *nonrotating origin* (NRO), which, as origin for right ascensions on the instantaneous equator, is now called the CIO; denoted as σ in Fig. 2.22).

Rather than successive transformations involving precessional elements and nutation angles, the transformation is more direct in terms of the coordinates, d and E . The additional parameter s defines the instantaneous origin of right ascension as an NRO (see below). Analogous to (2.51) and (2.55),

$$\begin{aligned} \mathbf{r} &= \mathbf{R}_3(-s)\mathbf{R}_3(-E)\mathbf{R}_2(d)\mathbf{R}_3(E)\mathbf{r}_0 \\ &= \mathbf{Q}^\top \mathbf{r}_0, \end{aligned} \quad (2.58)$$

which is easily derived by considering the successive rotations as the origin point transforms from the reference origin, Σ_0 , to the instantaneous origin, σ (Fig. 2.22). Equation (2.58) not only replaces (2.56), but also incorporates the new conventions for defining the intermediate origin in right ascension (it is no longer the true vernal equinox). The IERS Conventions 2003 (and later) define the transformation matrix, \mathbf{Q} , as a rotation from the system of the instantaneous pole and origin to the reference system.

The total rotation rate of the pole, P , in inertial space is due to the rates in the coordinates, d, E , and in the

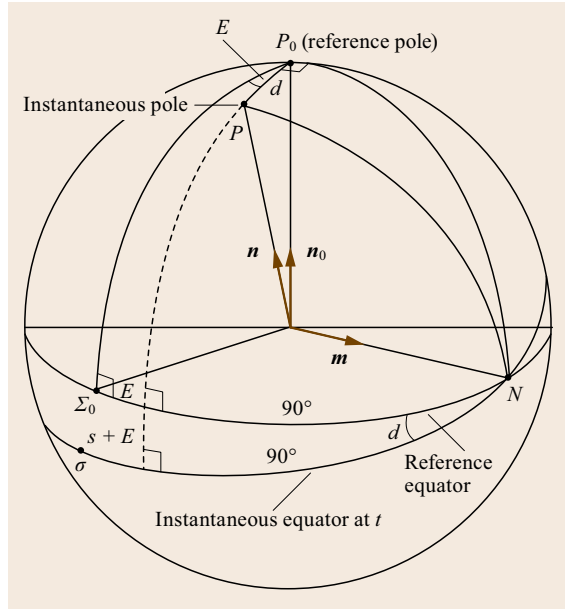


Fig. 2.22 Coordinates of the instantaneous pole in the celestial reference system

parameter, s . Defining three noncolinear unit vectors, $\mathbf{n}_0, \mathbf{m}, \mathbf{n}$, as shown in Fig. 2.22, the total rotation rate may be expressed as

$$\boldsymbol{\Theta} = \mathbf{n}_0 \dot{E} + \mathbf{m} \dot{d} - \mathbf{n} (\dot{E} + \dot{s}), \quad (2.59)$$

where the dots denote time derivatives. Now, s is chosen so that the total rotation rate, $\boldsymbol{\Theta}$, has no component along \mathbf{n} . That is, s defines the origin point, σ , on the instantaneous equator that has no rotation rate about the corresponding polar axis (it is thus a nonrotating origin). This condition is formulated as $\boldsymbol{\Theta} \cdot \mathbf{n} = 0$, meaning that there is no component of the total rotation rate along the instantaneous polar axis. Since $\mathbf{n} \cdot \mathbf{m} = 0$ and $\mathbf{n} \cdot \mathbf{n}_0 = \cos d$, (2.59) implies that

$$\dot{s} = (\cos d - 1) \dot{E}. \quad (2.60)$$

Defining coordinates

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} \sin d \cos E \\ \sin d \sin E \\ \cos d \end{pmatrix}, \quad (2.61)$$

it is easily shown that

$$\mathbf{Q} = \begin{pmatrix} 1 - aX^2 & -aXY & X \\ -aXY & 1 - aY^2 & Y \\ -X & -Y & 1 - a(X^2 + Y^2) \end{pmatrix} \mathbf{R}_3(s) \quad (2.62)$$

where $a = 1/(1 + \cos d)$. Furthermore, since

$$X\dot{Y} - Y\dot{X} = -\dot{E}(Z^2 - 1), \quad (2.63)$$

(2.60) integrates to

$$s = s_0 - \int_{t_0}^t \frac{X\dot{Y} - Y\dot{X}}{1 + Z} dt, \quad (2.64)$$

where $s_0 = s(t_0)$ is chosen so as to ensure continuity with the previous definition of the origin point at the epoch January 1, 2003.

Expressions for X and Y can be obtained directly from the precession and nutation equations [2.86]. For the latest IAU 2006/2000A precession–nutation models [2.6],

$$\begin{aligned} X = & -0.016617'' + 2004.191898''\tau \\ & - 0.4297829''\tau^2 - 0.19861834''\tau^3 \\ & - 0.000007578''\tau^4 - 0.0000059285''\tau^5 \\ & + \sum_{i=1}^n (e_i \sin A_i + e'_i \cos A_i) \\ Y = & -0.006951'' - 0.025896''\tau \\ & - 22.4072747''\tau^2 + 0.00190059''\tau^3 \\ & + 0.001112526''\tau^4 - 0.0000001358''\tau^5 \\ & + \sum_{i=1}^n (f_i \sin A_i + f'_i \cos A_i), \end{aligned} \quad (2.65)$$

where τ is given by (2.46), the coefficients, e_i, e'_i, f_i, f'_i are polynomials in τ , and the angles, A_i , are given by (2.54) including, for the higher indices, i , expressions for the longitudes of the planets (see Tables 5.2a,b in the electronic supplement [2.84] to the IERS Conventions 2010 [2.6]).

The corresponding series expression for the parameter s includes all terms larger than $0.5 \mu\text{as}$ (micro-arcsec), as well as the constant s_0

$$\begin{aligned} s = & -\frac{1}{2}XY + 94 + 3808.65\tau \\ & - 122.68\tau^2 - 72574.11\tau^3 \\ & + \sum_k C_k \sin \alpha_k + \sum_k D_k \sin \beta_k \\ & + \sum_k E_k \tau \cos \gamma_k + \sum_k F_k \tau^2 \sin \theta_k (\mu\text{as}). \end{aligned} \quad (2.66)$$

The coefficients C_k, D_k, E_k, F_k and the arguments, $\alpha_k, \beta_k, \gamma_k, \theta_k$, are elaborated by [2.6, p. 59]. Values of s

are less than $0.01''$ (until the early 2030s) and can be ignored for transformations at that level of accuracy.

The transformation formulas (2.65) and (2.66) yield an accuracy of about $0.3 \cdot 10^{-3}''$ in the position of the pole and incorporate the frame bias described in Sect. 2.4.

2.5.3 Polar Motion

The previous sections describe Earth's orientation from the celestial perspective – how the direction of an axis, such as the spin axis, progresses in time on the celestial sphere due to precession and nutation. From the terrestrial view, however, the spin axis and various other axes associated with Earth's rotation and geometry also exhibit motion with respect to the Earth's crust due to the natural dynamics of the rotation. Euler's equations describe the motion of the principal (geometric) axes for a rigid body, but because the Earth is partially fluid and elastic, the motion of these axes is not accurately predictable.

The details of the theoretical and mathematical developments of the dynamics equations for elastic rotating bodies may be found in [2.37]. These dynamics are influenced both by the internal composition and fluid characteristics of the Earth (nonforced, or free motion) and external gravitational torques that deform the Earth (forced motion). For example, the free motion of the principal axis (also called *figure axis*) corresponding to Earth's polar axis of symmetry has a circular diurnal motion relative a mean fixed location (*mean Tisserand axis*) with radius of about 60 m. The spin and angular momentum axes, on the other hand, have an order of magnitude smaller motion due to their greater stability, or relative insensitivity to Earth's deformation.

The change in direction of an axis, such as the instantaneous spin axis, of the Earth with respect to the surface of the Earth is called *polar motion* (also *wobble*). The motion is described by local coordinates, x_p, y_p , with respect to the reference pole of the terrestrial reference system. Figure 2.23 shows the polar motion coordinates for the CIP; note the defined direction of y , which is opposite to the y -axis of the right-handed system of Fig. 2.7. Viewed as horizontal Cartesian coordinates, their values change by only a few meters over several years; typically they are given by the subtended central angle, where $1''$ corresponds approximately to 30 m on the Earth's surface.

The principal component of polar motion is the *Chandler wobble*. This is basically the free Eulerian motion which would have a period of about 304 days, based on the moments of inertia of the Earth, if the Earth were a rigid body. Due to the elastic yielding of the Earth, resulting in displacements of the maximum

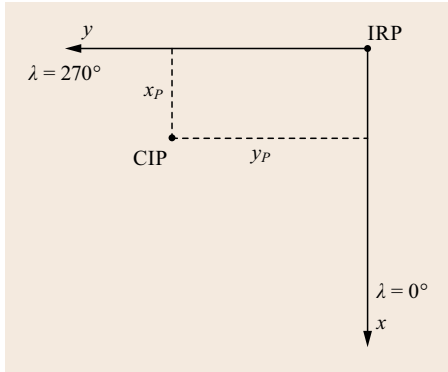


Fig. 2.23 Polar motion coordinates

moment of inertia, this motion has a longer period of about 430 days. S. C. Chandler observed and analyzed this discrepancy in the period in 1891; and, Newcomb gave the dynamical explanation [2.79, p. 80], thus also proving that the Earth is, in fact, not a rigid body. The period of this main component of polar motion is called the *Chandler period*; its amplitude is about 0.2 arcsec. Other components of polar motion include the approximately annual signal due to the redistribution of masses by way of meteorological and geophysical processes, with an amplitude of about 0.05–0.1", and the *nearly diurnal free wobble*, due to the slight misalignments of the rotation axes of the mantle and liquid outer core (also known as the *free core nutation*, with magnitude of about $0.1\text{--}0.3 \cdot 10^{-3}$ " and period of about 430 days with respect to the celestial sphere). Finally, there is the so-called *polar wander*, which is the secular motion of the pole. During 1900–2000, Earth's spin axis wandered about 0.004" per year in the direction of the 280° meridian. Figure 2.24 shows the polar motion for the period 2000–2010, and also the general drift for the last 110 years.

If \mathbf{r}_e is a unit vector that defines a geocentric direction of a point in the terrestrial reference system in terms of spherical coordinates

$$\mathbf{r}_e = \begin{pmatrix} \cos \lambda \cos \phi \\ \sin \lambda \cos \phi \\ \sin \phi \end{pmatrix}, \quad (2.67)$$

then the transformation from the terrestrial reference pole to the instantaneous, or intermediate pole (CIP), is given with appropriate rotations by

$$\begin{aligned} \mathbf{r}_i &= \mathbf{R}_1(y_p)\mathbf{R}_2(x_p)\mathbf{r}_e \\ &= \mathbf{W}\mathbf{r}_e \end{aligned} \quad (2.68)$$

Just as the instantaneous celestial system has a nonrotating origin for right ascension, one may define an

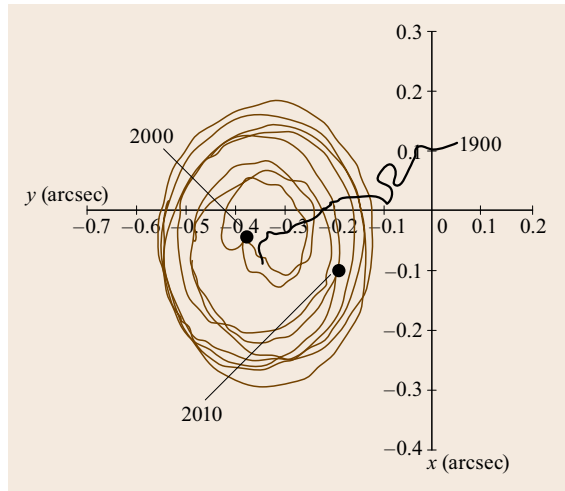


Fig. 2.24 Polar motion from 2000 to 2010, and polar wander since 1900. Polar motion coordinates are obtained from IERS and are smoothed to obtain the trend

instantaneous terrestrial system that has a nonrotating origin for longitudes, called the *Terrestrial Intermediate Origin (TIO)*. In this way, the only difference between the instantaneous celestial and terrestrial systems is Earth's rotation; the polar axes are the same.

With a derivation completely analogous to that for the precession–nutation matrix, \mathbf{Q} , the polar motion matrix is

$$\begin{aligned} \mathbf{W} &= \mathbf{R}_3(-s') \\ &\times \begin{pmatrix} 1 - a'x_p^2 & a'x_py_p & -x_p \\ a'x_py_p & 1 - a'y_p^2 & y_p \\ x_p & -y_p & 1 - a'(x_p^2 + y_p^2) \end{pmatrix}, \end{aligned} \quad (2.69)$$

where $a' = 1/2 + (x_p^2 + y_p^2)/8$. The parameter s' defines the location of the TIO on the instantaneous equator through an expression that is analogous to (2.66). By neglecting terms of second and higher orders, the exact equation (2.69) is approximately equal to

$$\mathbf{W} = \mathbf{R}_3(-s')\mathbf{R}_1(y_p)\mathbf{R}_2(x_p). \quad (2.70)$$

Furthermore, s' is significant only because of the largest components of polar motion and an approximate model is given by

$$s' = -0.0015'' \left(\frac{a_c^2}{1.2} + a_a^2 \right) \tau, \quad (2.71)$$

where a_c and a_a are the amplitudes, in arcsec, of the Chandler wobble ($\mathcal{O}(0.2''$) and the annual wobble

($\mathcal{O}(0.05'')$). Hence, the magnitude of s' is of the order of $0.1 \cdot 10^{-3}$ arcsec.

The polar motion coordinates are tabulated by the IERS as part of the Earth Orientation Parameters (EOP) and predicted on the basis of observations, such as from VLBI and satellite ranging. Thus, \mathbf{W} is a function of time, but there are no analytic models for polar motion as there are for precession and nutation. For the highest precision, the polar motion coordinates should be amended to include motions corresponding to nutations and tidal effects [2.6, Chaps. 5 and 8] with periods less than 2 days in the GCRS in order to comply with the definition of the intermediate pole.

2.5.4 Transformations

It is the current convention to formulate the transformation between celestial and terrestrial reference systems via an intermediate system. This intermediate, or true, or epoch-of-date system describes either precession and nutation when transformed from the celestial reference system, or polar motion and Earth rotation when transformed from the terrestrial reference system. As a dynamical system it is not a reference system since coordinates in this system vary significantly in time. For this reason, the intermediate system has also been called an *ephemeris* system. The newer *intermediate* nomenclature, more descriptive of the system's function, was adopted through a number of resolutions by the IAU during 2000–2006.

The ideal choice of the intermediate system largely falls on the choice of polar axis since the choice for the origin of the intermediate right ascension is now fixed by the nonrotating origin. In 1979 this pole was defined as having no motions with periods less than 2 days either with respect to the celestial or the terrestrial reference systems. The 2-day restriction on periods was consistent with the observational capability at the time to resolve such motions. The *Celestial Ephemeris Pole* (CEP), thus defined, divided the observable polar motion and predictable precession/nutations.

With improved VLBI observational techniques and data processing, shorter periods of motion could be discerned and in 2000 the IAU resolved to refine the definition of the intermediate pole. The newly named *celestial intermediate pole* (CIP) by definition, like the CEP, moves on the celestial sphere with periods greater than 2 days (frequencies less than ± 0.5 cycles per sidereal day, cpsd). This includes all the conventional predictable precessions and nutations produced by external gravitational torques on the Earth. Also included are the observed polar motions within ± 0.5 cpsd of Earth's diurnal rotation frequency (the *diurnal retrograde band*) since it can be shown that they are equivalent to nu-

tations with periods larger than 2 days. The terrestrial motions of the CIP, on the other hand, are defined to be those with frequencies outside the diurnal retrograde band. They not only include the conventional polar motions, such as the Chandler wobble, but also the high-frequency nutations, which are equivalent to polar motions outside this band. For additional details on these conventions, see [2.78, 88] and [2.75, p. 86].

It has been argued [2.89] that the intermediate pole is not essential in the transformation between the terrestrial and the celestial frame and that a combination of model and observations in a single transformation avoids much confusion and debate about the definition of the CIP. However, with current conventions the practical transformation between celestial and terrestrial reference frames combines the transformations (2.58) and (2.68) with Earth rotation,

$$\mathbf{r}_{\text{TRS}} = \mathbf{W}^T \mathbf{R}_3(\theta) \mathbf{Q}^T \mathbf{r}_{\text{CRS}}, \quad (2.72)$$

where θ is the Earth rotation angle (Sect. 2.1.3). This is called the *CIO method* of transformation, referring to the new convention of defining the origin for right ascension in the intermediate celestial system by the nonrotating origin. Alternatively, the so-called *equinox method*, uses the Greenwich sidereal time for the angle of Earth's rotation and the traditional precession and nutation series, given by (2.56)

$$\mathbf{r}_{\text{TRS}} = \mathbf{W}^T \mathbf{R}_3(\text{GAST}) \mathbf{NPB} \mathbf{r}_{\text{CRS}}. \quad (2.73)$$

where a small rotation, \mathbf{B} , is included to account for the frame bias.

Equations (2.72) and (2.73), of course, can be reversed to obtain coordinates in the CRF from coordinates in the terrestrial reference frame by noting that each rotation matrix is orthogonal – its inverse is its transpose

$$\mathbf{r}_{\text{CRS}} = \mathbf{Q} \mathbf{R}_3^T(\theta) \mathbf{W} \mathbf{r}_{\text{TRS}} \quad (2.74)$$

and

$$\mathbf{r}_{\text{CRS}} = \mathbf{B}^T \mathbf{P}^T \mathbf{N}^T \mathbf{R}_3^T(\text{GAST}) \mathbf{W} \mathbf{r}_{\text{TRS}}. \quad (2.75)$$

In applying the transformation (2.72), or (2.73), to observed points on the celestial sphere, it is important to consider any observational effects on the celestial coordinates of objects, such as actual, or proper motion (e.g., of stars), parallax due to the observer's changing position relative to the barycenter, and aberration due to the velocity of the observer relative to the barycenter. These effects are of primary interest for directional (e.g., optical or VLBI) observations of celestial bodies

but of limited relevance for GNSS data processing. For a detailed description, interested readers are referred to [2.5].

No matter whether the *equinox method* or the *CIO method* is adopted, the CRF-to-TRF transformation is characterized by extremely lengthy series expansions of the respective rotation angles. In order to facilitate the correct and consistent application of the conventional transformation, all relevant coefficients are made available in electronic form [2.84] by the IERS. Furthermore, various computer implementations of the transformations (or selected aspects thereof) are of-

fered by the IAU, the IERS, and individual authors. Such software may be applied directly, as a reference for validating independent implementations, or simply for better understanding of the underlying transformation concepts. Common examples include, for example, the IAU Standards of Fundamental Astronomy (SOFA, [2.90]) and the AstroRef package of [2.74]. Computational and implementation issues of the transformations are addressed in [2.91, 92]. Among others, the authors highlight the benefit of interpolating from a grid of pre-computed values, when evaluating the transformation for a dense set of nearby epoch values.

2.6 Perspectives

This chapter has introduced the basic concepts of modern space–time reference systems and frames that have jointly been developed by astronomers and geodesists as a basis of their work. They enable a concise description of the Earth’s motion in space and the location of objects on or near the Earth. Users of global navigation satellite systems are inevitably confronted with the issue of coordinates and reference systems, when it comes to the exchange and proper understanding of measured positions. Since GNSS provides essentially four-dimensional observations, with time as the fourth component of the navigation solution, the underlying concepts and conventions of time measurements are therefore equally important in all aspects of GNSS navigation.

Different GNSSs such as GPS, GLONASS, BeiDou, and Galileo have historically employed different time frames (realized by independent atomic clocks) and spatial reference frames (realized by different fundamental reference stations and partly different techniques). This affects the satellite orbit and clock information provided to the users and impacts a consistent navigation solution based on observations of multiple GNSS constellations. Fortunately, much progress has been achieved over the past decade. Individual frame realizations as used by the various GNSSs today exhibit centimeter-level differences that are well below the typical meter level accuracy of broadcast navigation information. Still, however, systematic time offsets (at the 10–100 ns level) between GNSS-specific time scales need to be carefully considered in the positioning and taken into account in the employed algorithms (Chap. 21).

Considering the high-level of accuracy that can today be achieved through carrier-phase-based GNSS positioning techniques, users are confronted with the fact that the Earth’s crust is far from solid and itself subject to permanent changes. This includes both long-

term changes such as tectonic plate motion (Sect. 2.3.2) but also periodic site shifts due to solid Earth and ocean tides (Sect. 2.3.5). Even though differential GNSS positioning techniques (Chap. 26) can offer (relative) accuracies down to the millimeter level, their use is largely unaffected by such intricate details. Undifferentiated, precise point positioning (PPP) techniques, in contrast, aim at providing absolute positions in a global reference frame. Here, a proper understanding of the underlying frame definitions and the consistent application of conventional corrections (e.g., for frame motion or tides) in the PPP software becomes an essential aspect of the GNSS data processing (Chap. 25). Similarity transformations between different regional and global frames (Sect. 2.3.4) or the transition between ellipsoidal and geoid heights (Sect. 2.3.1) are likewise important aspects of GNSS surveying (Chap. 35) and geodesy (Chap. 36).

While most precise GNSS users can confine themselves to a proper understanding of terrestrial reference systems and frames, the relation between celestial and terrestrial frames as discussed in Sect. 2.5 is likewise of relevance for various specific aspects of GNSS data processing. This includes, for example, the generation of GNSS precise orbit products (Chap. 34) and the GNSS-based precise orbit determination of satellites in low Earth orbit (Chap. 32). Satellite orbits and their equations of motion are most naturally expressed in a celestial frame, while the locations of GNSS monitoring stations are best described in a terrestrial frame. Conventional relations for the CRF-to-TRF transformation are essential for consistency of products obtained by individual analysis centers. On the other hand, the joint adjustment of satellite orbits, site locations, and Earth orientation parameters has become a vital part of space geodesy (Chap. 36) and contributes to a continued improvement of reference frames and the understanding of Earth rotation.

References

- 2.1 D.D. McCarthy, K.P. Seidelmann: *Time: From Earth Rotation to Atomic Physics* (Wiley-VCH, Weinheim 2009)
- 2.2 K. Lambeck: *Geophysical Geodesy, The Slow Deformations of the Earth* (Clarendon, Oxford 1988)
- 2.3 B.N. Taylor, A. Thompson (Eds.): *The International System of Units (SI)*, NIST SP 330 (National Institute of Standards and Technology, Gaithersburg 2008)
- 2.4 SI Brochure: The International System of Units (SI), 8th edn. (Bureau International des Poids et Mesures, Paris 2006)
- 2.5 P.K. Seidelmann: *Explanatory Supplement to the Astronomical Almanac* (Univ. Science Books, Mill Valley 1992)
- 2.6 G. Petit, B. Luzum: *IERS Conventions*, IERS Technical Note No. 36 (Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt 2010)
- 2.7 C. Audoin, B. Guinot: *The Measurement of Time: Time, Frequency and the Atomic Clock* (Cambridge Univ. Press, Cambridge 2001)
- 2.8 Bureau International des Poids et Mesures: BIPM Circular T, <ftp://ftp2.bipm.org/pub/tai/publication/cirt>
- 2.9 SI Brochure: Practical realization of the definition of the unit of time. In: *The International System of Units (SI)*, 8th edn. (Bureau International des Poids et Mesures, Paris, 2006) App. 2
- 2.10 D.D. McCarthy: Using UTC to determine the Earth's rotation angle, Proc. Coll. Explor. Implic. Redefining Coord. Univers. Time (UTC), Exton, ed. by S.L. Allen, J.H. Seago, R.L. Seaman (Univelt, San Diego 2011) pp. 105–116
- 2.11 Standard-Frequency and Time-Signal Emissions, ITU-R Recommendation TF.460–6 (International Telecommunication Union, Radio-communication Bureau, Geneva, Feb. 2002)
- 2.12 USNO: TAI minus UTC time difference <ftp://maia.usno.navy.mil/ser7/tai-utc.dat>
- 2.13 R.A. Nelson, D.D. McCarthy: S.N. Malys, J. Levine, B. Guinot, H. F. Fliegel, R. L. Beard, T. R. Bartholomew: The leap second: its history and possible future, *Metrologia* **38**(6), 509 (2001)
- 2.14 D. Finkleman, J.H. Seago, P.K. Seidelmann: The debate over UTC and leap seconds, Proc. AIAA Guid. Navig. Control Conf. Toronto (AIAA, Reston 2010), AIAA 2010–8391
- 2.15 P.K. Seidelmann, J.H. Seago: Time scales, their users, and leap seconds, *Metrologia* **48**(4), S186–S194 (2011)
- 2.16 W. Lewandowski, E.F. Arias: GNSS Times and UTC, *Metrologia* **48**(4), S219–S224 (2011)
- 2.17 K.R. Brown Jr.: The theory of the GPS composite clock, Proc. ION GPS 91, Albuquerque (ION, Virginia 1991) pp. 223–241
- 2.18 A.L. Satin, W.A. Feess, H.F. Fliegel, C.H. Yinger: GPS composite clock software performance, Proc. 22rd Annu. PTI Meet. Vienna (JPL, Pasadena 1991) pp. 529–546
- 2.19 H.S. Mobbs, S.T. Hutsell: Refining monitor station weighting in the GPS composite clock, Proc. 29th Annu. PTI Meet. Long Beach (JPL, Pasadena 1997)
- 2.20 Navstar GPS Space Segment/Navigation User Segment Interfaces, Interface Specification, IS-GPS-200H, 24 Sep. 2013 (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo 2013)
- 2.21 T.E. Parker, D. Matsakis: Time and frequency dissemination: Advances in GPS transfer techniques, *GPS World* **15**(11), 32–38 (2004)
- 2.22 Global Navigation Satellite System GLONASS – Interface Control Document, v5.1, (Russian Institute of Space Device Engineering, Moscow 2008)
- 2.23 P. Zhang, C. Xu, C. Hu, Y. Chen, J. Zhao: Time scales and time transformations among satellite navigation systems, Proc. CSNC 2012, Guanzhou, Vol. II, ed. by J. Sun, J. Liu, Y. Yang, S. Fan (Springer, Berlin 2012) pp. 491–502
- 2.24 A.V. Druzhin, V. Palchikov: Current state and perspectives of UTC(SU) broadcast by GLONASS, 9th Meet. Int. Comm. GNSS (ICG), Prague (UNOOSA, Vienna 2014) pp. 1–9
- 2.25 R. Zanello, M. Mascarello, L. Galleani, P. Tavella, E. Detoma, A. Bellotti: The Galileo precise timing facility, Proc. IEEE FCS 2007 21st EFTF, Geneva (2007) pp. 458–462
- 2.26 X. Stehlin, Q. Wang, F. Jeanneret, P. Rochat, E. Detoma: Galileo system time physical generation, Proc. 38th Annu. PTI Meet. Washington, DC (JPL, Pasadena 2006) pp. 395–406
- 2.27 C. Han, Y. Yang, Z. Cai: BeiDou navigation satellite system and its time scales, *Metrologia* **48**(4), S213–S218 (2011)
- 2.28 R. Hlaváč, M. Löscher, F. Luongo, J. Hahn: Timing infrastructure for Galileo system, Proc. 20th EFTF, Braunschweig (2006) pp. 391–398
- 2.29 BeiDou Navigation Satellite System Signal In Space Interface Control Document – Open Service Signal, Version 2.0 (China Satellite Navigation Office, 2013)
- 2.30 W. Torge, J. Müller: *Geodesy* (Walter de Gruyter, Berlin 2012)
- 2.31 K.M. Borkowski: Accurate algorithms to transform geocentric to geodetic coordinates, *Bull. Géodésique* **63**(1), 50–56 (1989)
- 2.32 D.D. McCarthy: *IERS Conventions (1996)*, IERS Technical Note No. 21 (Observatoire de Paris, Paris 1996)
- 2.33 T. Fukushima: Transformation from Cartesian to geodetic coordinates accelerated by Halley's method, *J. Geod.* **79**(12), 689–693 (2006)
- 2.34 H. Moritz: Geodetic reference system 1980, *Bull. Géodésique* **54**(3), 395–405 (1980)
- 2.35 E. Groten: Fundamental parameters and current (2004) best estimates of the parameters of common relevance to astronomy, geodesy, and geodynamics, *J. Geod.* **77**, 724–731 (2004)
- 2.36 J. Kovalevsky, I.I. Mueller: Comments on conventional terrestrial and quasi-inertial reference systems. In: *Reference Coordinate Systems for Earth Dynamics*, ed. by E.M. Gaposchkin, B. Kołczek (D. Reidel, Dordrecht 1981) pp. 375–384
- 2.37 H. Moritz, I.I. Mueller: *Earth Rotation: Theory and Observation* (Unger, New York 1987)

- 2.38 Geodetic Glossary (National Geodetic Survey, National Oceanic and Atmospheric Administration, Rockville 1986)
- 2.39 G. Seeber: *Satellite Geodesy: Foundations, Methods, and Applications* (Walter de Gruyter, Berlin 2003)
- 2.40 H. Schuh, D. Behrend: VLBI: A fascinating technique for geodesy and astrometry, *J. Geodyn.* **61**, 68–80 (2012)
- 2.41 R.A. Snay, T. Soler: Continuously operating reference station (CORS): History, applications, and future enhancements, *J. Surv. Eng.* **134**(4), 95–104 (2008)
- 2.42 B. Hofmann-Wellenhof, H. Moritz: *Physical Geodesy* (Springer, Berlin 2005)
- 2.43 N.K. Pavlis, S.A. Holmes, S.C. Kenyon, J.K. Factor: The development and evaluation of the Earth Gravitational Model 2008 (EGM2008), *J. Geophys. Res. Solid Earth* **117**(B4), 2156–2202 (2012)
- 2.44 Standardization Agreement Navstar Global Positioning System (GPS) System Characteristics, STANAG 4294, 1st edn. (North Atlantic Treaty Organization, 1993)
- 2.45 S. Malys, J.H. Seago, N.K. Pavlis, P.K. Seidelmann, G.H. Kaplan: Why the Greenwich meridian moved, *J. Geod.* **89**(12), 1263–1272 (2015)
- 2.46 Z. Altamimi, X. Collilieux, L. Métivier: ITRF2008: An improved solution of the International Terrestrial Reference Frame, *J. Geod.* **85**(8), 457–473 (2011)
- 2.47 M.R. Pearlman, J.J. Degnan, J.M. Bosworth: The international laser ranging service, *Adv. Space Res.* **30**(2), 135–143 (2002)
- 2.48 L. Combrinck: Satellite laser ranging. In: *Sciences of Geodesy*, Vol. I, ed. by G. Xu (Springer, Berlin 2010) pp. 301–338
- 2.49 M. Meindl, G. Beutler, D. Thaller, R. Dach, A. Jäggi: Geocenter coordinates estimated from GNSS data as viewed by perturbation theory, *Adv. Space Res.* **51**(7), 1047–1064 (2013)
- 2.50 S.P. Kuzin, S.K. Tatevian, S.G. Valeev, V.A. Fashutdinova: Studies of the geocenter motion using 16-years DORIS data, *Adv. Space Res.* **46**(10), 1292–1298 (2010)
- 2.51 Z. Altamimi, C. Boucher, P. Sillard: New trends for the realization of the international terrestrial reference system, *Adv. Space Res.* **30**(2), 175–184 (2002)
- 2.52 Z. Altamimi, P. Sillard, C. Boucher: ITRF2000: A new release of the International Terrestrial Reference Frame for Earth science applications, *J. Geophys. Res.* **107**(B10), 2214 (2002)
- 2.53 D.F. Argus, R.G. Gordon: No-net-rotation model of current plate velocities incorporating plate motion model NUVEL-1, *Geophys. Res. Lett.* **18**(11), 2039–2042 (1991)
- 2.54 C. DeMets, R.G. Gordon, D.F. Argus, S. Stein: Effect of recent revisions to the geomagnetic reversal time scale on estimates of current plate motions, *Geophys. Res. Lett.* **21**(20), 2191–2194 (1994)
- 2.55 Z. Altamimi, L. Métivier, X. Collilieux: ITRF2008 plate motion model, *J. Geophys. Res.* **117**(B07402), 1–14 (2012)
- 2.56 Department of Defense World Geodetic System 1984 (WGS84): Its definition and relationships with local geodetic systems, Publication NIMA TR8350.2, 3rd edn., amendm. 1 (National Imagery and Mapping Agency, 2000)
- 2.57 Supplement to Department of Defense World Geodetic System 1984 Technical Report, Part I, DMA TR 8350.2–A (Defense Mapping Agency, Washington 1987)
- 2.58 M.J. Merrigan, E.R. Swift, R.F. Wong, J.T. Saffel: A refinement to the World Geodetic System 1984 reference frame, *Proc. ION GPS 2002*, Portland (IOM, Virginia 2002) pp. 1519–1529
- 2.59 R.F. Wong, C.M. Rollins, C.F. Minter: Recent Updates to the WGS 84 Reference Frame, *Proc. ION GNSS 2012*, Nashv. (ION, Virginia 2012) pp. 1164–1172
- 2.60 S.G. Revnivykh: GLONASS status and progress, 47th CGSIC Meet. Fort Worth (2007)
- 2.61 Parametry Zemli 1990 (PZ–90.11) Reference document (Military Topographic Department of the General Staff of Armed Forces of the Russian Federation, Moscow 2014)
- 2.62 A.N. Zueva, E.V. Novikov, D.I. Pleshakov, I.V. Gusev: System of geodetic parameters “Parametry Zemli 1990” PZ–90.11, 9th Meet. Int. Comm. GNSS (ICG), Work. Group D, Prague (UNOOSA, Vienna 2014)
- 2.63 Y. Yang: Chinese Geodetic Coordinate System 2000, *Chin. Sci. Bull.* **54**(15), 2714–2721 (2009)
- 2.64 G. Gendt, Z. Altamimi, R. Dach, W. Söhne, T. Springer, GGSP Prototype Team: GGSP: Realisation and maintenance of the Galileo terrestrial reference frame, *Adv. Space Res.* **47**(2), 174–185 (2011)
- 2.65 D. D. McCarthy: *IERS Conventions (1992)*, IERS Technical Note No. 13 (Observatoire de Paris, Paris 1992)
- 2.66 International Terrestrial Reference Frame: ITRF Transformation Parameters, ITRF Website http://itrf.ensg.ign.fr/trans_para.php
- 2.67 T. Soler, R.A. Snay: Transforming positions and velocities between the International Terrestrial Reference Frame of 2000 and North American Datum of 1983, *J. Surv. Eng.* **130**(2), 49–55 (2004)
- 2.68 IAG: IAG resolutions adopted at the XXth IUGG General Assembly in Vienna, *Bulletin Géodésique* **66**(2), 132–133 (1992)
- 2.69 M. Poutanen, M. Vermeer, J. Mäkinen: The permanent tide in GPS positioning, *J. Geod.* **70**(8), 499–504 (1996)
- 2.70 P.M. Mathews, B.A. Buffett, I.I. Shapiro: Love numbers for a rotating spheroidal Earth: New definitions and numerical values, *Geophys. Res. Lett.* **22**(5), 579–582 (1995)
- 2.71 J.M. Wahr: Deformation induced by polar motion, *J. Geophys. Res. Solid Earth* **90**(B11), 9363–9368 (1985)
- 2.72 P.M. Mathews, B.A. Buffett, I.I. Shapiro: Love numbers for diurnal tides: Relation to wobble admittances and resonance expansions, *J. Geophys. Res.* **100**(B6), 9935–9948 (1995)
- 2.73 W.E. Farrell: Deformation of the Earth by surface loads, *Rev. Geophys.* **10**(3), 761–797 (1972)
- 2.74 M. Soffel, R. Langhans: *Space-Time Reference Systems* (Springer, Berlin 2012)
- 2.75 V. Dehant, P.M. Mathews: *Precession, Nutation and Wobble of the Earth* (Cambridge Univ. Press, Cambridge 2015)

- 2.76 E.W. Woollard: *Theory of the Rotation of the Earth Around its Center of Mass*, Astronomical Papers Vol. XV Part 1 (U.S. Naval Observatory, Washington, D.C. 1953)
- 2.77 J.H. Lieske, T. Lederle, W. Fricke, B. Morando: Expressions for the precession quantities based upon the IAU/1976/system of astronomical constants, *Astron. Astrophys.* **58**, 1–16 (1977)
- 2.78 N. Capitaine, P.T. Wallace, J. Chapront: Expressions for IAU 2000 precession quantities, *Astron. & Astrophys.* **412**(2), 567–586 (2003)
- 2.79 I.I. Mueller: *Spherical and Practical Astronomy as Applied to Geodesy* (F. Ungar, New York 1969)
- 2.80 H. Goldstein, C.P. Poole, J.L. Safko: *Classical Mechanics* (Addison Wesley, San Francisco 2000)
- 2.81 N. Capitaine, P.T. Wallace, J. Chapront: Improvement of the IAU 2000 precession model, *Astron. & Astrophys.* **432**(1), 355–367 (2005)
- 2.82 J.P. Vinti: *Orbital and Celestial Mechanics* (AIAA, Reston 1998)
- 2.83 H. Kinoshita: Theory of the rotation of the rigid Earth, *Celest. Mech.* **15**(3), 277–326 (1977)
- 2.84 International Earth Rotation and Reference Systems Service: IERS Conventions 2010, electronic supplement http://62.161.69.134/iers/conv2010/conv2010_c5.html
- 2.85 N. Capitaine, B. Guinot, J. Souchay: A non-rotating origin on the instantaneous equator: Definition, properties and use, *Celest. Mech.* **39**(3), 283–307 (1986)
- 2.86 N. Capitaine: The celestial pole coordinates, *Celest. Mech. Dyn. Astron.* **48**(2), 127–143 (1990)
- 2.87 D.D. McCarthy, G. Petit: *IERS Conventions (2003)*, IERS Technical Note No. 36 (Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt 2004)
- 2.88 P.M. Mathews, P. Bretagnon: Polar motions equivalent to high frequency nutations for a nonrigid Earth with anelastic mantle, *Astron. & Astrophys.* **400**(3), 1113–1128 (2003)
- 2.89 P. Mathews, T. Herring: On the reference pole for Earth orientation and UT1, *Proc. IAU Colloq. 180: Towards Models Constants Sub-Microarcsecond Astrom.* Washington DC, ed. by K.J. Johnston, D.D. McCarthy, B.J. Luzum, G.H. Kaplan (US Naval Observatory, Washington, DC 2000) pp. 164–170
- 2.90 IAU SOFA Board: IAU SOFA Software Collection (International Astronomical Union), IAU SOFA Center <http://www.iausofa.org>
- 2.91 D.A. Vallado, J.H. Seago, P.K. Seidelmann: Implementation issues surrounding the new IAU reference systems for astrodynamics, *Proc. 16th AAS/AIAA Space Flight Mech. Conf. Tampa (AAS, San Diego 2006)*, pp. 1–22, AAS 06–134
- 2.92 V. Coppola, J.H. Seago, D.A. Vallado: The IAU 2000A and IAU 2006 precession–nutations theories and their implementation, *Proc. 19th AAS/AIAA Space Flight Mech. Meet. Savannah (AAS, San Diego 2009)*, pp. 1–20, AAS 09–159

Satellite Orbits

3. Satellite Orbits and Attitude

Urs Hugentobler, Oliver Montenbruck

This chapter discusses fundamentals of orbital dynamics and provides a description of key perturbations affecting global navigation satellite system (GNSS) satellites along with their impact on the orbits. Models for perturbing accelerations including Earth gravity, third body perturbations, surface forces, and relativistic corrections are described with emphasis on empirical and semiempirical solar radiation pressure models. Long-term evolution of GNSS orbits and orbit keeping maneuvers are discussed. The concepts of broadcast orbit models such as almanac models, analytical ephemeris models and numerical ephemeris models used by current GNSS systems are presented along with cook book algorithms and a summary of their performance. Complementary to the discussion of GNSS satellite orbits, the chapter introduces the basic concepts of GNSS satellite attitude, which are, for example, required to describe the antenna location relative to the center-of-mass.

3.1 Keplerian Motion	59
3.1.1 Basic Properties	59
3.1.2 Keplerian Orbit Model	61
3.1.3 Ground Track and Visibility	63
3.2 Orbit Perturbations	66
3.2.1 Orbit Representation	66
3.2.2 Perturbing Accelerations	67
3.2.3 Perturbations at GNSS Satellite Altitude	71
3.2.4 Radiation Pressure	72
3.2.5 Long-Term Evolution	74
3.2.6 Orbit Accuracy	77
3.3 Broadcast Orbit Models	79
3.3.1 Almanac Models	80
3.3.2 Keplerian Ephemeris Models	81
3.3.3 Cartesian Ephemeris Model	83
3.3.4 Broadcast Ephemeris Generation and Performance	83
3.4 Attitude	85
References	87

3.1 Keplerian Motion

Long before the launch of the first man-made satellite, the motion of planets and moons in the solar system has been a subject of extensive research. Celestial mechanics [3.1, 2] has helped us to understand the properties of orbital motion around a central body under the action of gravitational forces from first principles. It allowed to accurately predict planetary and lunar motion for the benefit of science but likewise for timekeeping and navigation.

3.1.1 Basic Properties

The fundamental properties of planetary orbits were originally identified by Johannes Kepler based on his careful analysis of Tycho Brahe's observations of planet Mars. Applied to the case of an Earth-orbiting satellite, Kepler's three laws of planetary motion describe

the following properties:

1. The orbit of a satellite is an ellipse with the Earth located in one of its foci.
2. The radius vector covers a constant area per unit time interval (law of areas).
3. The square of the orbital period increases with the third power of the mean distance from the center of the Earth.

In the most general sense, satellite orbits are described by conic sections, which also allows for parabolic and hyperbolic trajectories. Such orbits may be attained by spacecraft on their way to or from the solar system. However, ellipses represent the only type of closed and periodic orbits around the central body (Fig. 3.1).

The shape of an ellipse is determined by the two radii a (semimajor axis) and b (semiminor axis) along

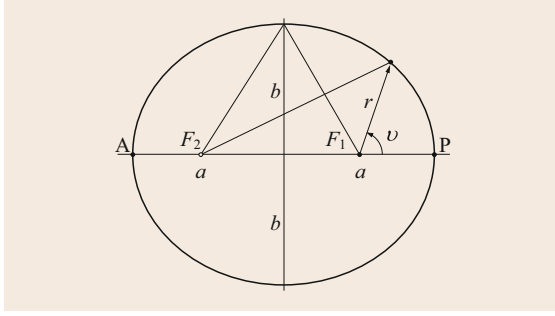


Fig. 3.1 An ellipse describes the set of all points for which the sum of the distances from the two given foci (F_1, F_2) has a constant value. In the case of a planetary satellite orbit, the central body is located in one of these focal points. The line of apsides passes through both foci and connects the apocenter (A) and the pericenter (P) of the orbit

the principal axes, or, equivalently the eccentricity

$$e = \frac{\sqrt{a^2 - b^2}}{a} \quad (3.1)$$

The distance from the focus varies between a minimum of $r_{\min} = a(1 - e)$ at pericenter (P) and a maximum of $r_{\max} = a(1 + e)$ at apocenter (A). Denoting by ν the angle between the radius vector and the pericenter direction, the radius is described by the conic section equation

$$r = \frac{a(1 - e^2)}{1 + e \cos \nu} \quad (3.2)$$

which likewise holds for ellipses, parabolas, and hyperbolas [3.3].

Kepler's laws are rigorously valid for the so-called two-body problem, which addresses the motion of two point-masses under the action of their mutual gravitational attraction. Additional forces such as the gravitational attraction of third bodies, the gravitational field of nonspherical, extended bodies, or diverse types of non-gravitational forces may induce perturbations, which are further addressed in Sect. 3.2 and discussed in full detail in common textbooks of astrodynamics [3.4, 5]. Still, Keplerian motion remains an important concept for understanding and describing the trajectories of Earth orbiting satellites.

As can be shown, Kepler's laws can be fully explained from Newton's law of gravity [3.2, 5]. They are a direct consequence of the fact that the gravitational attraction is always direct along the line joining the two bodies and decreases with the second power of the distance.

Denoting the position vector and distance of a satellite with respect to the center of the Earth by \mathbf{r} and

$r = ||\mathbf{r}||$, respectively, the corresponding acceleration is given by

$$\ddot{\mathbf{r}} = -\frac{GM_{\oplus}}{r^2} \frac{\mathbf{r}}{r} \quad (3.3)$$

Here, $G \approx 6.674 \cdot 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$ [3.6] is the universal constant of gravity and $M_{\oplus} \approx 5.973 \cdot 10^{24} \text{ kg}$ denotes the mass of the Earth. The latter is about 20 orders of magnitude larger than the satellite's mass m , which has therefore been neglected in Eq. 3.3. It may be noted that the gravitational acceleration (3.3) depends only on the product of the gravitational constant and the Earth mass. While the individual values are only known to roughly four digits, the product

$$GM_{\oplus} = 3.986004415 \cdot 10^{14} \text{ m}^3 \text{ s}^{-2} \quad (3.4)$$

can in fact be determined with very high precision from the analysis of satellite orbits [3.7]. Note that the number given in (3.4) is consistent with the terrestrial scale defined by the rate of clocks at the rotating geoid.

As can be recognized from the unit vector \mathbf{r}/r in (3.3), the gravitational acceleration is always directed radially inward. Changes of the velocity are thus confined to a plane spanned by the instantaneous position and velocity vectors. In the absence of perturbing accelerations directed perpendicular to \mathbf{r} and $\dot{\mathbf{r}}$, the satellite will thus move in a constant orbital plane at all times.

Physically speaking, the angular momentum vector $\mathbf{h} = \mathbf{r} \times \dot{\mathbf{r}}$ does not change over time, since

$$\dot{\mathbf{h}} = \mathbf{r} \times \ddot{\mathbf{r}} + \dot{\mathbf{r}} \times \dot{\mathbf{r}} = \mathbf{0} \quad (3.5)$$

whenever the acceleration is directed along the radius vector. As illustrated in Fig. 3.2, the modulus of the angular momentum vector matches twice the area ΔA swept by the satellite's radius vector within a time interval Δt

$$h = ||\mathbf{r} \times \dot{\mathbf{r}}|| = 2 \frac{\Delta S}{\Delta t} \quad (3.6)$$

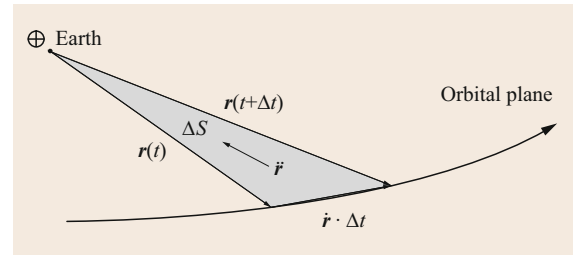


Fig. 3.2 A central acceleration $\ddot{\mathbf{r}}$ does not alter the plane of the satellite's orbit and results in a constant areal velocity $\Delta S/\Delta t$ of the radius vector

The areal velocity is thus constant, which proves the validity of Kepler's second law.

While the law of areas is equally valid for all types of central forces, the property of an elliptic motion is specific to an attractive force varying with the inverse square of the distance. In this case, the so-called *Runge–Lenz* or *Laplace* vector

$$\mathbf{A} = -\mathbf{h} \times \dot{\mathbf{r}} - GM_{\oplus} \frac{\mathbf{r}}{r} \quad (3.7)$$

is found to be constant throughout the orbit [3.8]. By forming the scalar product with the radius vector, the conic section equation (3.2) with angle $\nu = \angle(\mathbf{A}, \mathbf{r})$, eccentricity $e = \|\mathbf{A}\|/GM_{\oplus}$ and parameter $a(1 - e^2) = h^2/GM_{\oplus}$ can finally be obtained.

Kepler's third law, while valid for all eccentricities, can most easily be justified for a circular orbit. Considering a satellite orbit of radius $r = a$ with orbital period T and angular velocity $n = 2\pi/T$, the equality of centrifugal and gravitational acceleration yields the relation

$$a^3 n^2 = GM_{\oplus}. \quad (3.8)$$

Among others, this representation of Kepler's third law is useful to find the semimajor axis at which a given orbital period is achieved. For navigation satellites, a certain repeat rate of the orbit relative to the Earth is typically desired and the orbital period is therefore chosen to match a rationale multiple of the Earth's sidereal rotation period of 23 h 56 min (Table 3.1).

3.1.2 Keplerian Orbit Model

The time dependence of the orbital motion is fully determined by the law of areas and may be derived from the geometric properties of the ellipse. Both the angular velocity and the orbital velocity are highest at perigee

(the point nearest to the Earth) and lowest at apogee (the most distant point of the orbit). For a mathematical description, it is common to introduce an auxiliary quantity termed the eccentric anomaly E (Fig. 3.3).

Making use of this value, the perifocal coordinates, that is, the position of the satellite relative to the central body and a Cartesian coordinate system oriented in the pericenter direction, may be expressed as

$$\begin{aligned} x_p &= r \cos \nu = a (\cos E - e), \\ y_p &= r \sin \nu = a \sqrt{1 - e^2} \sin E. \end{aligned} \quad (3.9)$$

Furthermore, the shaded area encompassed by the radius vector, the orbit arc and the line of apsides (Fig. 3.3) is given by

$$S(E) = \frac{1}{2} a^2 \sqrt{1 - e^2} (E - e \sin E). \quad (3.10)$$

Using (3.6) for computing the area swept by the radius vector between time t_0 when the satellite passes the pericenter and the satellite position at time t , we finally get

$$E - e \sin E = M = n(t - t_0). \quad (3.11)$$

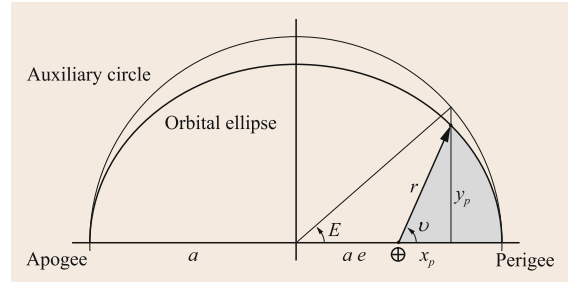


Fig. 3.3 Geometric relation of the eccentric anomaly E and the true anomaly ν in an elliptic orbit

Table 3.1 Representative orbital parameters (period, semimajor axis a , height h , eccentricity e , and inclination i) of global and regional navigation satellite systems satellites

System	Period (rev/d _{sid})	Period (h)	a (km)	h (km)	e	i (°)
GLONASS	17/8	11 h 16 min	25 510	19 130	0.0	64.8
GPS	2/1	11 h 58 min	26 560	20 180	0.0	55
BeiDou (MEO)	13/7	12 h 53 min	27 910	21 530	0.0	55
Galileo	17/10	14 h 05 min	29 600	23 220	0.0	56
QZSS	1/1	23 h 56 min	42 160	35 790	0.1	43
BeiDou IGSO	1/1	23 h 56 min	42 160	35 790	0.0	55
NavIC (IGSO)	1/1	23 h 56 min	42 160	35 790	0.0	27
BeiDou/NavIC/QZSS GEO; SBAS	1/1	23 h 56 min	42 160	35 790	0.0	≤ 2

GLONASS: Globalnaja Nawigazionnaja Sputnikowaja Sistema (Russian Global Navigation Satellite System); GPS: Global Positioning System; MEO: medium altitude Earth orbit; QZSS: Quasi-Zenith Satellite System; IGSO: inclined geo-synchronous orbit; NavIC: Navigation with Indian Constellation; GEO: geostationary Earth orbit; SBAS: satellite-based augmentation system

Here, M denotes the mean anomaly which grows linearly with time at the mean motion. Equation (3.11) links geometric quantities with time and is commonly known as Kepler's equation. Unfortunately, it cannot be used directly to compute the eccentric anomaly $E(t)$ at a given time t , but must be solved in an iterative manner. A first approximation $E_0 = M$ is typically given by the mean anomaly itself. Refined values can then be obtained using Newton's method

$$E_{i+1} = E_i - \frac{E_i - e \sin E_i - M}{1 - e \cos E_i}. \quad (3.12)$$

Except for QZSS, which adopts an intentional eccentricity of $e \approx 0.1$, the orbits of most navigation satellites are almost circular with eccentricities $e = 0.01$ or less. Given the quadratic convergence of Newton's method, a few iterations are therefore sufficient to compute the eccentric anomaly E with an accuracy of 10 digits, which provides a subcentimeter-level accuracy in the resulting positions.

The spacecraft velocity in the perifocal system is obtained from differentiation of (3.9) and (3.11), which results in

$$\begin{aligned} \dot{x}_p &= -\frac{\sqrt{GM_\oplus a}}{r} \sin E, \\ \dot{y}_p &= +\frac{\sqrt{GM_\oplus a}}{r} \sqrt{1-e^2} \cos E. \end{aligned} \quad (3.13)$$

The description of the satellite orbit has so far been confined to the orbital plane and the line of apsides. In order to express the motion in a global reference frame, such as the International Celestial Reference Frame (ICRF; Chap. 2), the orientation of the orbit is commonly described by a set of three angles (Fig. 3.4):

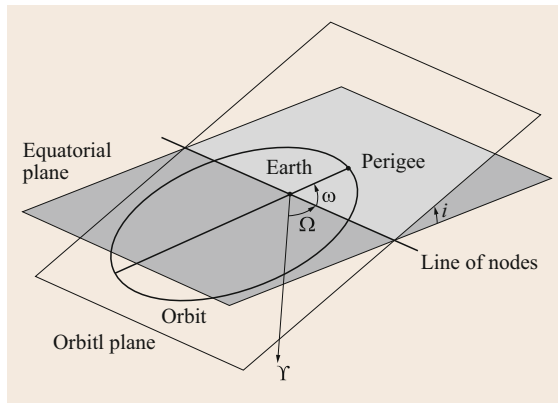


Fig. 3.4 Orbital elements defining the orientation of the orbital plane and the line of apsides

- The *right ascension of the ascending node* (RAAN, Ω) measures the angle between the x -direction of the celestial coordinate system (roughly aligned with the vernal equinox) and the ascending node. The latter describes the point of the orbit, in which the satellite crosses the equatorial plane from South to North (i. e., in *ascending* direction).
- The *inclination* i specifies the angle between the orbital plane and the reference plane (i. e., the celestial equator). More specifically, the inclination is defined as the angle between the North direction and the angular momentum vector of the orbit. Satellites orbiting the Earth in the direction of the Earth rotation (i. e., anticlockwise as seen from North) exhibit inclinations of $0^\circ \leq i < 90^\circ$. Those orbiting the Earth in the opposite (retrograde) direction are described by inclinations of $90^\circ < i \leq 180^\circ$.
- Finally, the *argument of perigee* ω measures the angle between the ascending node and the pericenter of the orbit, measured in the direction of satellite motion.

For global navigation satellite system (GNSS) satellites in medium altitude Earth orbits (MEOs), inclination values near 55° are most commonly adopted in an effort to achieve good visibility, moderate orbital perturbations, and acceptable launch cost. However, different values have been adopted for GLONASS as well as the various regional systems (Table 3.1). Geostationary satellites, such as those of the various space-based augmentation systems (SBAS; Chap. 12) apply a near-zero inclination (along with an orbital period of about 24 h) to maintain an almost constant position in the sky for terrestrial users.

Altogether, the three angles (Ω, i, ω) uniquely define the spatial orientation of the orbit and allow a transformation from the perifocal coordinates (x_p, y_p) to celestial coordinates $\mathbf{r}_{\text{ICRF}} = (x, y, z)_{\text{ICRF}}$ through a series of three consecutive rotations

$$\mathbf{r}_{\text{ICRF}} = \mathbf{R}_3(-\Omega) \mathbf{R}_1(-i) \mathbf{R}_3(-\omega) \begin{pmatrix} x_p \\ y_p \\ 0 \end{pmatrix}. \quad (3.14)$$

A corresponding transformation

$$\dot{\mathbf{r}}_{\text{ICRF}} = \mathbf{R}_3(-\Omega) \mathbf{R}_1(-i) \mathbf{R}_3(-\omega) \cdot \begin{pmatrix} \dot{x}_p \\ \dot{y}_p \\ 0 \end{pmatrix} \quad (3.15)$$

applies for the velocity, since the orientation of the orbital plane remains constant in the Keplerian orbit model. The matrices $\mathbf{R}_1(\phi)$ and $\mathbf{R}_3(\phi)$ used in the above

expression used in (3.14) and (3.15) describe elementary rotations about the x - and z -axis, respectively, and are defined as

$$\mathbf{R}_1(\phi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & +\cos \phi & +\sin \phi \\ 0 & -\sin \phi & +\cos \phi \end{pmatrix} \quad (3.16)$$

and

$$\mathbf{R}_3(\phi) = \begin{pmatrix} +\cos \phi & +\sin \phi & 0 \\ -\sin \phi & +\cos \phi & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (3.17)$$

Combining (3.9) with (3.14) and evaluating the individual transformations, the satellite position may finally be expressed as

$$\mathbf{r}_{\text{ICRF}} = r \begin{pmatrix} \cos u \cos \Omega - \sin u \cos i \sin \Omega \\ \cos u \sin \Omega + \sin u \cos i \cos \Omega \\ \sin u \sin i \end{pmatrix}. \quad (3.18)$$

Here,

$$u = \omega + \nu \quad (3.19)$$

denotes the instantaneous angle between the radius vector and the ascending node, which is also known as *argument of latitude*.

Alltogether, six quantities ($a, e, i, \Omega, \omega, M$) are used to uniquely define the position and velocity in the Keplerian orbit model at a given time. They are commonly termed the Keplerian orbital elements.

3.1.3 Ground Track and Visibility

While the satellite is moving around the Earth in its orbital plane, the Earth continuously rotates beneath the satellite. Neglecting, for simplicity, variations in the orientation of the Earth rotation axis (such as precession, nutation, and polar motion; Chap. 2), the transformation between the celestial reference frame (such as the ICRF) and the terrestrial reference frame (such as the International Terrestrial Reference Frame (ITRF)) is described by a uniform rotation about the common z -axis with angular velocity

$$\omega_{\oplus} \approx 7.292 \cdot 10^{-5} \text{ rad/s}. \quad (3.20)$$

The Earth rotation angle (or *mean sidereal time*) Θ is approximately given by

$$\Theta \approx 280.46^\circ + (360.985653^\circ/\text{d}) d, \quad (3.21)$$

where d denotes the number of days elapsed since 1 January 2000 Universal Time.

In the case of a nonrotating Earth, the satellite's footprint would describe a great circle on the surface of the Earth that covers a latitude range of $-i \leq \varphi \leq +i$ (Fig. 3.5) and repeats itself after each revolution. Due to the Earth's rotation, however, repeated equator crossings will no longer take place at the same geographic longitude but are shifted West by an angle $\Delta\lambda = \omega_{\oplus} T$ that matches the Earth rotation during the orbital period.

The resulting ground tracks are illustrated in Fig. 3.6 for representative satellites of the various GNSS systems. For MEO satellites of the GPS, GLONASS, BeiDou, and Galileo systems a periodic variation of latitude versus longitude is obtained. Depending on their orbital periods, consecutive ascending node crossings are offset in Eastern direction by angles between 148° (Galileo) and 191° (GLONASS). As a special case, a 180° offset is obtained for GPS, which results in a repetition of the ground track after just two revolutions. As discussed earlier, the orbital periods of GLONASS, BeiDou (MEO), and Galileo satellites are also rational fractions of the Earth rotation period (see second column in Table 3.1). Accordingly, their ground tracks likewise repeat after several revolutions. However, the repeat cycles are substantially longer than for GPS and last a total of eight, seven, and 10 days, respectively.

For inclined geosynchronous orbits (IGSOs), which are employed in QZSS, BeiDou, and NavIC, the ground track attains a distinct figure-of-eight. Since the orbital period of IGSO satellites matches that of the

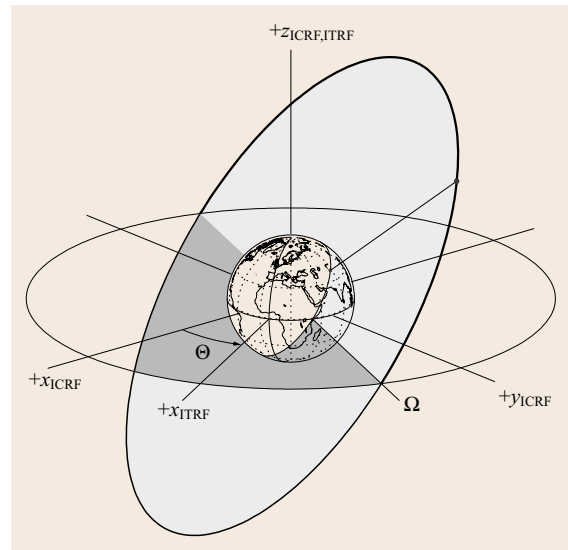


Fig. 3.5 GPS satellite orbit relative to the Earth

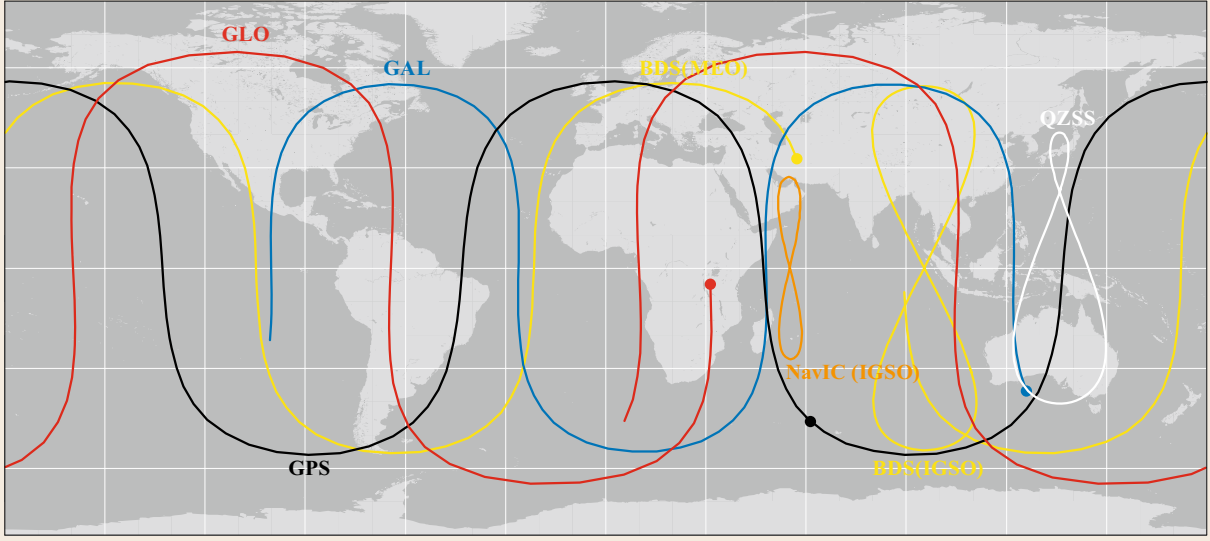


Fig. 3.6 Representative ground tracks of GNSS satellites in medium altitude Earth orbit (MEO) and IGSO over a 24 h period. The footprint of MEO satellites proceeds from West (left) to East (right). End points are marked by filled circles. The IGSO ground track describes a figure of eight, which is traversed in clockwise direction in the Northern Hemisphere and in counter-clockwise direction in the Southern Hemisphere

Earth rotation, all equator crossings take place at the same longitude. However, the projected angular velocity varies with the satellite's latitude and is less than that of the Earth at the ascending and descending nodes. As such, the ground track crosses the equator from East to West and the Northern part is traversed in a clockwise sense, while the Southern loop is traversed in a counter-clockwise direction.

Satellites in geostationary orbit (GEO), finally, employ both a near-zero inclination as well as an orbital period closely matching a sidereal day. As a result, they exhibit an essentially stationary footprint throughout a day at a selected point on the equator. In practice, small variations arise due to nonperfect conditions and orbital perturbations but these are typically confined to the order of 1° in longitude and latitude.

In accord with the orbital properties discussed earlier, GEO and IGSO satellites are only ever visible from certain regions of the Earth, while MEO satellite can be viewed from any location on Earth for at least part of their orbit.

To describe the motion of a GNSS satellite relative to a station on the surface of the Earth, a local tangential coordinate system originating in the station and aligned with the East (E), North (N), and Up (U) direction is most commonly employed. Considering a station at geographic longitude λ and latitude φ , the transformation from the terrestrial reference frame to the ENU frame is

described by the rotation matrix

$$\mathbf{E} = \begin{pmatrix} -\sin \lambda & +\cos \lambda & 0 \\ -\sin \varphi \cos \lambda & -\sin \varphi \sin \lambda & +\cos \varphi \\ +\cos \varphi \cos \lambda & +\cos \varphi \sin \lambda & +\sin \varphi \end{pmatrix}. \quad (3.22)$$

For a given station at position \mathbf{r}_{sta} to a GNSS satellite at position \mathbf{r}_{sat} (both expressed in the terrestrial frame) the line-of-sight (LOS) unit vector in the ENU frame is

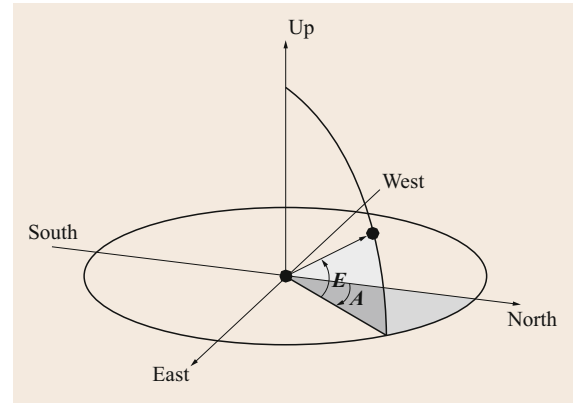


Fig. 3.7 Azimuth and elevation of a GNSS satellite in the local topocentric coordinate system of the observer

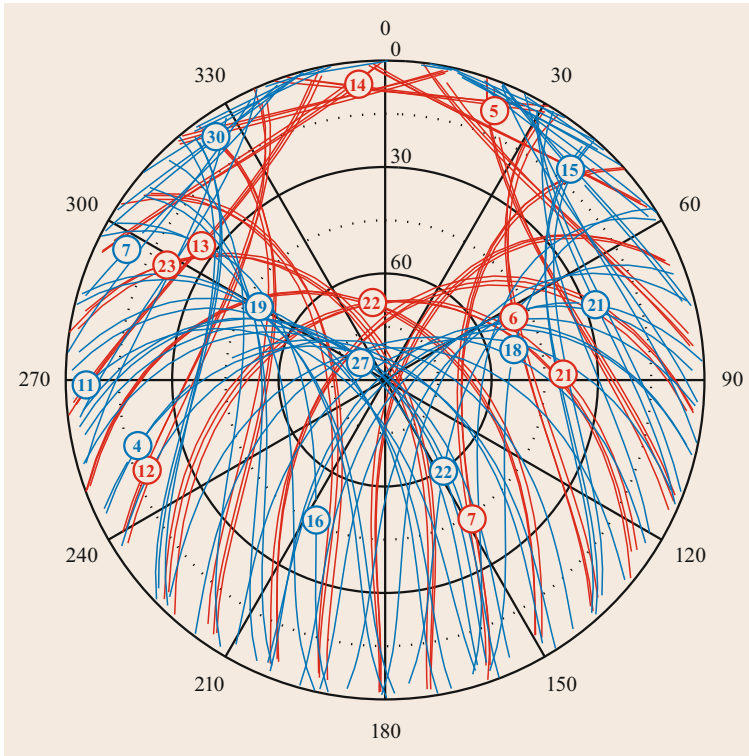


Fig. 3.8 Azimuth and elevation of GPS and GLONASS satellites for an observer in Munich ($\varphi = 48.8^\circ\text{N}$) over a 24 h period on 1st April 2015. At the midnight epoch a total of 11 GPS satellites and 9 GLONASS satellites were visible above a 5° elevation threshold

thus given by

$$\mathbf{e} = \mathbf{E} \cdot \frac{\mathbf{r}_{\text{sat}} - \mathbf{r}_{\text{sta}}}{\|\mathbf{r}_{\text{sat}} - \mathbf{r}_{\text{sta}}\|}. \quad (3.23)$$

The satellite's *azimuth* describes the angle between North direction and the projection of the line-of-sight vector on the local horizontal plane, while the *elevation* gives the angle between the LOS and the horizon (Fig. 3.7).

A skyplot illustrating the motion of GPS and GLONASS satellites for a station in the Northern Hemisphere over a 1-day period is shown in Fig. 3.8. Satellites remain above the horizon for periods of up to about six hours. While the satellites are roughly symmetrically distributed in the Eastern and Western Hemispheres, a pronounced asymmetry in the North-South distribution may be recognized. In particular, no

GPS satellites are ever visible in a cone of about 40° semidiameter around the celestial pole. In view of their higher inclination, the GLONASS satellites offer a better sky coverage with an exclusion zone of about 30° semidiameter.

For polar stations, GNSS satellites of the GPS, Galileo and BeiDou MEO constellations achieve peak elevations of about 45° , which results in an unfavorable vertical dilution of precision (VDOP; Chap. 1) and thus a degraded vertical positioning performance. Again the situation is somewhat improved for GLONASS satellites, which attain a maximum elevation of about 57° . A largely symmetric distribution of visible satellites and a good overall sky coverage, in contrast, are obtained for stations near the equator. Here, the exclusion zone is made up of two semicircles centered around the North and South points on the horizon.

3.2 Orbit Perturbations

Section 3.1 describes the Keplerian motion of a satellite about a massive parent body under the sole influence of the gravitational force originating from a central body. Kepler's laws and the equations describing the satellite motion are strictly valid under the assumption that the central body can be considered as a point mass or has a spherically symmetrical mass distribution. In real situations the central body (the Earth) has a complex mass distribution which is varying under the influence of tidal deformation, additional bodies like Sun and Moon exert their gravitational influence, radiation pressure impacts the satellite, and atmospheric drag dissipates orbit energy. These perturbing forces affect the motion of the satellite in a complex way.

To describe the motion of a satellite in this more realistic situation, the two-body equation of motion given in (3.3) has to be extended to include the additional perturbing accelerations and may read

$$\ddot{\mathbf{r}} = -\frac{GM_{\oplus}}{r^2} \frac{\mathbf{r}}{r} + \mathbf{a}(\mathbf{r}, \dot{\mathbf{r}}, t) . \quad (3.24)$$

The perturbing acceleration \mathbf{a} is, in general, a function of the satellite's position, of the satellite's velocity (e.g., in case of drag), and of time. As the largest perturbation is about 1000 times smaller than the central force we may indeed consider the additional acceleration \mathbf{a} as a perturbation, allowing for approximative analytical solutions of (3.24) (e.g., [3.9–11]). In order to reach the highest accuracy the perturbed equation of motion is however solved using methods of numerical integration.

3.2.1 Orbit Representation

In the presence of perturbations, the orbit of a satellite is no longer an ellipse (or more generally, a conic sections) as described in the previous section. However, because the perturbations are small compared to the central term, the orbit of a satellite can still be considered as an ellipse with constantly varying parameters. Due to perturbations, the orientation of the orbital plane in space is in general no longer stable (i. e., the orbital angular momentum \mathbf{h} is no longer conserved) but varying.

In each point $\mathbf{r}(t)$ along the perturbed trajectory of the satellite a best fitting ellipse can be defined. This so-called *osculating ellipse* is tangent to the satellite trajectory at position $\mathbf{r}(t)$ and both orbit representations have in this point the identical vectorial velocity $\dot{\mathbf{r}}$. The perturbed orbit of the satellite can be considered as the envelope of the osculating ellipses associated

with each orbit position $\mathbf{r}(t)$ (Fig. 3.9). The perturbed orbit can thus be represented by the orbital elements of these osculating ellipses. This bijective relation between position and velocity on one side and time dependent so-called *osculating elements* on the other side

$$\{\mathbf{r}(t), \dot{\mathbf{r}}(t)\} \leftrightarrow \{a(t), e(t), i(t), \Omega(t), \omega(t), t(t_0)\}$$

thus allows us to uniquely describe the perturbed trajectory by its osculating elements.

In order to assess the impact of perturbing forces the time variations of the osculating elements may be studied. Perturbations induce typical variations in osculating elements. These can be classified into short-periodic, long-periodic, and secular variations. Short-periodic perturbations in the osculating elements have typical periods that are equal to the satellite's revolution period or integer fractions thereof while long-periodic perturbations have periods of weeks to many years. Secular perturbations show no periodic patterns but a continuous increase or decrease of an osculating element. They are typically observed in the right ascension of ascending node Ω and in the argument of perigee ω . As short- and long-periodic perturbations are bounded by their amplitudes, secular perturbations in the osculating elements will after sufficiently long time always dominate all other perturbations. They are thus of par-

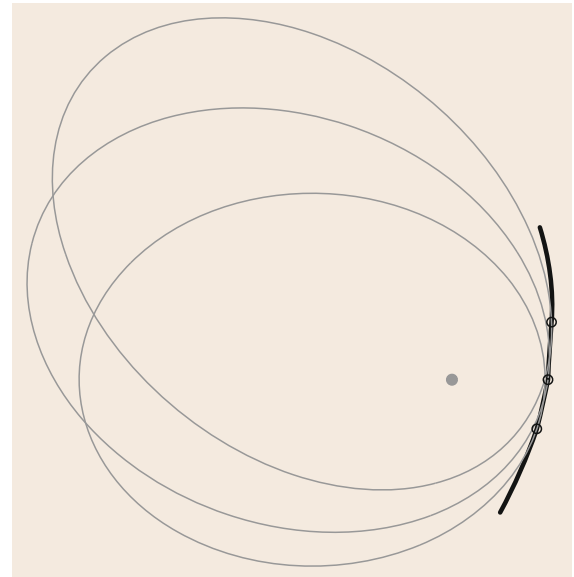


Fig. 3.9 True orbit (*bold*) approximated at successive positions by osculating ellipses (*thin lines*)

ticular interest when assessing the impact of perturbing accelerations on satellite orbits.

Figure 3.10 shows variations of four osculating elements of a Galileo satellite for a time period of 10 days. These perturbations in the orbital elements are induced by the strongest perturbing acceleration at GNSS altitude caused by the oblateness of the Earth. The osculating semimajor axis (Fig. 3.10a) shows short-periodic oscillations with a period of twice-per-revolution (7 h) and an amplitude of about 1.5 km. The eccentricity (Fig. 3.10b) shows short-periodic variations consisting of a superposition of once-per-revolution and three-times-per-revolution oscillations as well as a long-periodic perturbation caused by Sun and Moon. In the inclination (Fig. 3.10c) a long-periodic variation is superimposed to a short-periodic perturbations. The RAAN (Fig. 3.10d) finally shows a secular perturbation, a retrograde precession of -1.7 arcmin/day, superimposed by a short-periodic perturbation. Similar perturbations can be observed for the remaining orbital elements. They are also very similar for other GNSS satellites at MEO altitudes.

3.2.2 Perturbing Accelerations

For typical satellite orbits, the accelerations caused by the inhomogeneous mass distribution inside the Earth represent the strongest perturbations. The gravitational acceleration of the Earth may be described by the gradient ∇V of the gravitational potential $V(\mathbf{r})$ which may be represented in the space surrounding the Earth in terms of a spherical harmonic series in the form

$$V(\mathbf{r}) = \frac{GM_{\oplus}}{r} \sum_{n=0}^{\infty} \sum_{m=0}^n \left(\frac{R_{\oplus}}{r} \right)^n \times P_{nm}(\sin \varphi) (C_{nm} \cos m\lambda + S_{nm} \sin m\lambda), \quad (3.25)$$

where r , λ , and φ represent the spherical coordinates (radius, geocentric longitude, and latitude) of the position \mathbf{r} given in the Earth-fixed reference frame, R_{\oplus} represents the equatorial radius of the Earth, P_{nm} stands for the associated Legendre polynomials of degree n and order m , and C_{nm} and S_{nm} are the so-called Stokes coefficients [3.12]. Modern gravity fields obtained from

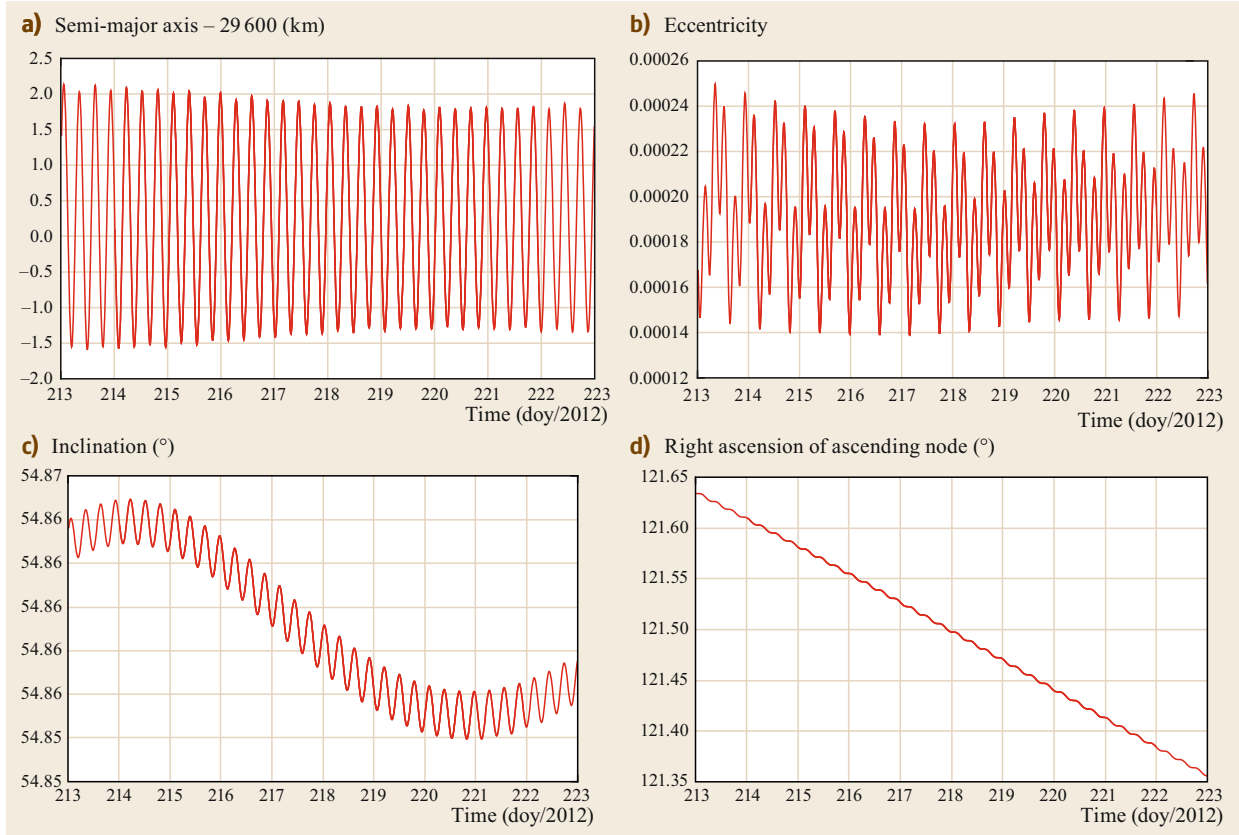


Fig. 3.10a–d Osculating elements for the Galileo IOV-1 satellite over 10 days from July 7 to August 10, 2012. Semimajor axis (a), offset by 29 600 km, eccentricity (b), inclination (c), right ascension of ascending node (d)

gravity field recovery missions are expanded up to high degree and order, for example, 280×280 for the static gravity field determined by the gravity field and steady state ocean circulation explorer (GOCE) [3.13], up to 2190×2190 for EGM2008 [3.14] which combines satellite information with terrestrial gravity measurements. Note that the perturbing acceleration exerted by gravity field terms of degree n decreases with a power of $n+2$ with the distance from the Earth's center. While for low orbiting satellites the gravity field may have to be considered up to degree and order 100, potential terms with degree above 8 cause perturbations below 10^{-11} m/s^2 at GNSS altitudes causing orbit errors at the sub-millimeter level and can be neglected with respect to other perturbations.

The first term $n=0$ in the harmonic expansion (3.25) corresponds to the gravitational potential of a spherically symmetric mass distribution which leads to the two-body acceleration, that is, the first term on the right-hand side of (3.24). As the gravity field is represented in the geocentric frame, the terms of degree $n=1$ vanish. The zonal term of degree 2, $C_{20} = -J_2 = -1.082 \cdot 10^{-3}$ reflects the flattening of the Earth. The net torque exerted by the perturbing acceleration resulting from the oblate mass distribution of the Earth causes a precession of the orbital planes in space as well as a rotation of the osculating ellipses within their planes. The secular part of this precession in the RAAN Ω and in the argument of perigee ω may be written in the following form

$$\dot{\Omega}|_{\text{secular}} \simeq -\frac{10.0^\circ/\text{d} \cos i}{\left(\frac{a}{R_\oplus}\right)^{7/2} (1-e^2)^2}, \quad (3.26)$$

$$\dot{\omega}|_{\text{secular}} \simeq +\frac{5.0^\circ/\text{d} (5 \cos^2 i - 1)}{\left(\frac{a}{R_\oplus}\right)^{7/2} (1-e^2)^2}. \quad (3.27)$$

We recognize that for orbital inclinations i smaller than 90° the precession of the node is retrograde (opposite to the direction of the Earth's rotation) while for inclinations larger than 90° (retrograde motion of the satellite) the precession of the node is prograde. A proper selection of the inclination $i > 90^\circ$ allows the exploitation of the oblateness perturbation for the establishment of sun-synchronous orbits that precess at the same rate as the mean motion of the Sun. Sun-synchronous orbits are particularly popular for Earth observation satellites at low Earth orbits but of no interest for GNSS satellites.

The equation for the secular precession of the perigee (3.27) shows that there is a critical inclination of $i_{\text{crit}} = 63.4^\circ$. The precession of the perigee is prograde resp. retrograde for inclinations smaller resp. larger than the critical value while the perigee shows no secu-

lar drift caused by oblateness if the inclination is equal to the critical value. It is interesting to note that the value of the critical inclination can be computed from the mathematical relation $5 \cos(i_{\text{crit}}) = 1$ and does not depend on the value of the zonal coefficient J_2 . Critical inclination is, e.g., exploited for Russian telecommunication satellites of the Molnya class. As GNSS orbits have typically low eccentricities the orbit inclinations are defined by other requirements than by perturbations in the location of the perigee.

Besides the Earth also other celestial bodies, in particular Sun and Moon, perturb the motion of Earth orbiting satellites. To model these third-body perturbations the Newtonian gravitational attraction is computed assuming point masses for the perturbing bodies. As the perturbing accelerations need to be represented in the geocentric frame, an additional term appears in the equation for the perturbing acceleration of third bodies

$$\mathbf{a}_i = -GM_i \left(\frac{\mathbf{r} - \mathbf{r}_i}{\|\mathbf{r} - \mathbf{r}_i\|^3} + \frac{\mathbf{r}_i}{\|\mathbf{r}_i\|^3} \right). \quad (3.28)$$

Here M_i represents the mass of the perturbing celestial body i , \mathbf{r}_i is the geocentric position vector of this body, and \mathbf{r} is the geocentric position vector of the perturbed satellite.

The second term on the right-hand side of (3.28) represents the (negative) perturbing acceleration of the celestial body acting on the Earth and accounts for the fact that the geocentric reference frame is not inertial but accelerated under the action of the perturbing celestial body. The gravitational third-body acceleration is represented in the geocentric frame by a so-called tidal acceleration which decreases with the third power of the distance of the perturbing body but increases linearly with the distance of the satellite from the center of the Earth. The third-body perturbations are also called direct tides.

In order to precisely model the motion of Earth satellites also the so-called indirect tides have to be considered. In fact, the gravitational forces from Sun and Moon cause a tidal deformation of the Earth's body. These solid Earth tides result in a time variable mass distribution of the Earth which results in its turn in a perturbation of the satellite orbits. Similarly the impact on satellite orbits from tidal mass variations in the oceans are considered. For the computation of the temporal variations of the Stokes coefficients in (3.25) due to solid Earth tides closed-form models are available from conventional models [3.7] while the temporal variations of the gravity field from ocean tides are computed from sophisticated ocean tide models that are based on hydrodynamic finite element models, (e.g., FES2004, [3.15]) or altimetry observations (e.g., EOT11a [3.16]).

As the equation of motion (3.24) is formulated in Euclidian space relativistic corrections are added as perturbing accelerations in order to account for the curvature of space–time. The main correction is the so-called Schwarzschild term which describes the space curvature caused by the mass of the Earth [3.7, 17]. The Schwarzschild perturbation is proportional to the mass of the Earth and decreases with the third power of the distance of the satellite from the center of the Earth. For low Earth satellites it is of the order of $2 \cdot 10^{-8} \text{ m/s}^2$ while it reaches $7 \cdot 10^{-11} \text{ m/s}^2$ for geosynchronous orbits. The Schwarzschild correction results in a negligible perigee precession of the order of 1 mas/d for GNSS satellites. As the Schwarzschild correction is an in-plane correction it does not affect the orientation of the orbital plane.

The Lense–Thirring effect – also called gravitomagnetic or frame dragging effect – is a general relativistic correction caused by the spinning mass of the Earth. It results in a precession of the reference frame which is modeled by Coriolis accelerations. The perturbing acceleration is proportional to the angular momentum of the Earth and decreases with the power $7/2$ of the distance of the satellite from the center of the Earth. The effect is of the order of $2 \cdot 10^{-10} \text{ m/s}^2$ for low Earth satellites and $4 \cdot 10^{-13} \text{ m/s}^2$ for geosynchronous orbits. The corresponding orbit precession varies between $500 \text{ } \mu\text{as/d}$ and $2 \text{ } \mu\text{as/d}$ from low to high satellites.

The de-Sitter or geodetic precession is caused by the space–time curvature induced by the mass of the Sun along the orbit of the Earth around the Sun. It causes a precession of the geocentric reference frame with respect to the fixed stars of 19.2 mas/yr or $53 \text{ } \mu\text{as/d}$. The corresponding Coriolis acceleration is proportional to the mass of the Sun and decreases from $5 \cdot 10^{-11} \text{ m/s}^2$ to $2 \cdot 10^{-11} \text{ m/s}^2$ from low to high satellites.

In addition to the gravitational accelerations discussed earlier the motion of satellites is also affected by nongravitational forces, also called surface forces. Contrary to gravitational forces, surface forces do not act on the mass of the satellite but on the surface of the satellite body. They are caused by interaction of particles or radiation with the outer satellite surfaces, for example, by atoms and ions in the high atmosphere causing drag forces, or by photons from the Sun reflected and absorbed by the illuminated surfaces and transferring momentum to the satellite thus causing a perturbing force. Characteristic for surface accelerations is their direct dependency on the satellite cross-section A and inverse dependency on the satellite mass m . A discussion may be found, for example, in [3.5, 18].

Atmospheric drag acceleration is caused by atmospheric particles at satellite orbit height and is propor-

tional to the air mass density. As air density decreases exponentially with height above the Earth’s surface, air drag is relevant for satellites below an altitude of about 2000 km. For GNSS satellites air drag is irrelevant. The simplest model of drag acceleration reads

$$\mathbf{a}_{\text{drag}} = -\frac{1}{2} C_D \frac{A}{m} \rho(\mathbf{r}) v_{\text{rel}}^2 \frac{\mathbf{v}_{\text{rel}}}{v_{\text{rel}}} . \quad (3.29)$$

The acceleration acts in the opposite direction to the velocity vector \mathbf{v}_{rel} relative to the atmosphere and is proportional to the square of the velocity for the laminar flow of air at satellite altitudes. The drag coefficient C_D depends on the aerodynamic properties of the satellite body and on the details of the interaction of the satellite surfaces and the air particles. Typical values are between 2 and 3. The air density $\rho(\mathbf{r})$ is very difficult to model. It depends on the thermospheric temperature which in its turn depends on the solar and geomagnetic activity. The drag acceleration at an altitude of 450 km is of the order of $1 \text{ } \mu\text{m/s}^2$ and increases to some $100 \text{ } \mu\text{m/s}^2$ at an altitude of 250 km. Depending on solar activity, the drag acceleration at given altitude may vary by more than one order of magnitude.

Radiation pressure is caused by the interaction of light with the surface of the satellite which results in a momentum transfer. For a satellite at position \mathbf{r} the simplest *cannon-ball* form for the acceleration from radiation pressure due to direct sunlight reads

$$\mathbf{a}_{\text{rpr}} = -\gamma C_R \frac{A}{m} \frac{S_0}{c} \left(\frac{1 \text{ AU}}{r_\odot} \right)^2 \frac{\mathbf{r}_\odot - \mathbf{r}}{\|\mathbf{r}_\odot - \mathbf{r}\|} . \quad (3.30)$$

where \mathbf{r}_\odot is the geocentric position vector of the Sun. S_0 represents the solar radiation flux of 1361 Wm^{-2} [3.19] at a distance of one astronomical unit which gives, divided by the speed of light c the radiation pressure of $4.539 \cdot 10^{-6} \text{ Nm}^{-2}$. The factor $(1 \text{ AU}/r_\odot)^2$ scales this flux to the current distance of the Earth from the Sun on its slightly eccentric orbit. C_R is the radiation pressure coefficient which depends on the shape and surface properties such as reflectivity and absorption, and γ represents the eclipse factor which is 1 in full sunlight and 0 in the shadow of the Earth or the Moon. More complex models consider the satellite structure decomposed into individual surfaces with specified size, orientation, and optical properties (*box-wing model*). The force caused by radiation pressure is then computed for each illuminated surface and added up to the total force. The accuracy of the model depends on the information provided by the manufacturer and on the known attitude of the satellite with respect to the Sun. Typical accelerations due to solar radiation are of the order of $1 \cdot 10^{-7} \text{ ms}^{-2}$.

Table 3.2 Average acceleration and orbit error after two revolutions for different orbit types and perturbations

Perturbation	GPS			Galileo			IGSO		
	Average acceleration (m/s ²)	Orbit error after 2 rev.		Average acceleration (m/s ²)	Orbit error after 2 rev.		Average acceleration (m/s ²)	Orbit error after 2 rev.	
		Initial conditions Fixed (m)	Adjusted (m)		Initial conditions Fixed (m)	Adjusted (m)		Initial conditions Fixed (m)	Adjusted (m)
Earth oblateness	5.7 · 10 ⁻⁵	23 000	3000	3.8 · 10 ⁻⁵	22 000	2700	9.1 · 10 ⁻⁶	16 000	1900
Direct tides Moon	3.0 · 10 ⁻⁶	1900	170	3.3 · 10 ⁻⁶	2700	270	4.7 · 10 ⁻⁶	12 000	1100
Direct tides Sun	1.6 · 10 ⁻⁶	930	90	1.7 · 10 ⁻⁶	1700	110	2.5 · 10 ⁻⁶	6900	480
Higher potential terms	3.7 · 10 ⁻⁷	360	32	2.4 · 10 ⁻⁷	340	30	5.6 · 10 ⁻⁸	1100	85
Direct solar rad.press.	1.0 · 10 ⁻⁷	220	32	1.0 · 10 ⁻⁷	290	44	1.0 · 10 ⁻⁷	860	130
Earth albedo	9.8 · 10 ⁻¹⁰	1.1	0.050	1.4 · 10 ⁻⁹	2.2	0.11	7.0 · 10 ⁻¹⁰	3.1	0.15
Solid Earth tides	1.1 · 10 ⁻⁹	0.70	0.044	7.4 · 10 ⁻¹⁰	0.67	0.034	1.8 · 10 ⁻¹⁰	0.47	0.024
Antenna thrust (100 W)	3.1 · 10 ⁻¹⁰	0.37	0.005	4.9 · 10 ⁻¹⁰	0.79	0.010	4.9 · 10 ⁻¹⁰	2.3	0.030
General relativity	2.8 · 10 ⁻¹⁰	0.33	0.004	2.1 · 10 ⁻¹⁰	0.33	0.004	7.1 · 10 ⁻¹¹	0.33	0.004
Venus (inf. conj.)	1.7 · 10 ⁻¹⁰	0.11	0.010	1.9 · 10 ⁻¹⁰	0.20	0.011	2.8 · 10 ⁻¹⁰	0.83	0.046
Ocean tides	1.2 · 10 ⁻¹⁰	0.10	0.009	7.5 · 10 ⁻¹¹	0.09	0.010	1.8 · 10 ⁻¹¹	0.13	0.009
Jupiter (opposit.)	2.3 · 10 ⁻¹¹	0.014	0.0014	2.5 · 10 ⁻¹¹	0.024	0.0018	3.6 · 10 ⁻¹¹	0.099	0.007
Pot. terms degree > 8	9.1 · 10 ⁻¹²	0.0054	0.0006	2.8 · 10 ⁻¹²	0.0022	0.0004	5.5 · 10 ⁻¹⁴	0.0009	0.0003
Mars (opposit.)	1.6 · 10 ⁻¹²	0.0011	0.0004	1.7 · 10 ⁻¹²	0.0016	0.0004	2.5 · 10 ⁻¹²	0.0067	0.0008

Apart from direct solar radiation also indirect radiation reflected from the Earth surface (albedo radiation) as well as infrared radiation emitted by the Earth represents an important fraction of the total radiation-induced perturbing acceleration for low orbiting satellites. Albedo models decompose the surface of the Earth into segments and compute for each of them the radiation impact on the satellite [3.20, 21]. Radiation absorbed by the satellite is re-radiated as thermal radiation causing an acceleration opposite to the direction of emission. For a thorough modeling of the effect a thermal model of the satellite has to be considered.

If the satellite is spinning delayed re-radiated thermal emission causes additional effects such as the Yarkovsky–Rubincam effect [3.22] caused by thermal radiation from the Earth and the Yarkovsky–Schach effect [3.23] caused by heating of the satellite by direct solar radiation. For a satellite like the geodetic laser geodynamic satellite (LAGEOS) these effects cause an alongtrack acceleration of the order of a few 10^{-12} ms^{-2} [3.24]. Finally, also the radio emission broadcast by a satellite causes a recoil force of the order of L/c where L is the emitted radio power and c is the speed of light. For an emission power of 100 W, the resulting force is about $3 \cdot 10^{-7} \text{ N}$.

The impact of solar wind can be computed from flux, velocity, and mass of the high-energetic particles ejected by the Sun and captured by the Earth's magnetosphere. For typical strong Solar events the resulting perturbing acceleration is some 4 orders of magnitude below that caused by solar radiation and can thus be neglected.

3.2.3 Perturbations at GNSS Satellite Altitude

After the general introduction to perturbations in Sect. 3.2.2, let us discuss the orbit perturbations for GNSS satellites. Table 3.2 summarizes the magnitude of perturbations and their impact on GPS and Galileo satellites as well as on IGSO satellites. For the latter, a near circular orbit with a semimajor axis of 42 164 km and an inclination of 55° from the BeiDou constellation is taken as an example. The table gives the mean accelerations of selected perturbations as well as their impact on the orbits. The values listed in the table are average values, computed as the root mean square of the values obtained for individual satellites.

The impact of the perturbing acceleration on the orbits is obtained by numerically integrating the orbits of an ensemble of satellites with and without the respective perturbation switched on. The orbit differences after two satellite revolutions (i. e., 1 d for GPS and

2 d for IGSO) were then determined and averaged. The numerical integration was performed in two different ways. The values in the first column give the differences obtained if the orbit is propagated with the same initial conditions. The orbit differences thus show the plain effect of the perturbations on the orbits.

The values in the second columns of the table were obtained by adjusting the initial conditions in order to allow an optimum fit of the orbit to the case with and without the perturbation enabled. The orbit differences after two revolutions are then significantly smaller than with fixed initial conditions. Because initial conditions always have to be estimated as part of an orbit determination process, these values more realistically represent the observable impact of the perturbations on the orbits. Since, in general, additional parameters such as empirical parameters of a radiation pressure model (Sect. 3.2.4) are adjusted, even these values are in fact pessimistic estimates.

We observe that the Earth's oblateness has by far the largest impact on GNSS satellites, resulting in orbit fit errors at the kilometer level. Accelerations by Sun and Moon are the second-largest effect causing orbit deformations at the order of several hundred meters. Although the gravitational attraction of the Sun at the position of the satellites is larger than that of the Moon due to the much higher mass of the Sun, the tidal acceleration, that is, the gravitational acceleration represented in the geocentric frame is larger for the Moon due to the much lower distance. Terms in the geopotential with higher degree and order than the oblateness have an impact on GNSS orbits at a similar level as radiation pressure. As terms with a degree above 8 have a negligible impact, a very modest harmonic expansion of the geopotential to 8×8 is sufficient for modeling GNSS satellite orbits.

While variations of the geopotential due to deformations of the solid Earth by tides from Sun and Moon have an impact on GNSS orbits at the several centimeter level, the effect of ocean tides is nearly an order of magnitude smaller. Third body perturbations from other planets have effects on the orbit at the centimeter or subcentimeter level. Perturbations for Venus, Mars, and Jupiter are given for their closest distance to the Earth, that is, for the inferior conjunction for Venus, that is, for the inner planet between Sun and Earth, and the opposition, that is, closest approach for the outer planets.

Contrary to gravitational accelerations, the effects related to interaction of the satellites with radiation from the Sun and the Earth are difficult to model. As they cause large orbit perturbations at the several ten meter level their modeling is a challenge for reaching an orbit accuracy at the few centimeter level.

3.2.4 Radiation Pressure

Solar radiation pressure is the largest nongravitational perturbation acting on GNSS satellites. It is significantly more difficult to model than gravitational perturbations, because it depends on the details of the satellite structure, dimensions, optical surface properties, and attitude, that is, information which is often not publicly available for GNSS satellites. Different types of radiation pressure models were thus developed to cope with this perturbation that causes orbit errors at the few hundred meter level within a day.

Nominal orientation of GNSS satellites requests that the navigation antenna points to the center of the Earth while the solar panels are oriented perpendicularly to the direction of the Sun (Sect. 3.4). As a consequence, the Sun moves in the body-fixed xz -plane and may illuminate only three surfaces of the body while the satellite orbits around the Earth. These are the front panel, where the navigation antenna is mounted ($+z$ -panel), the panel on the backside of the satellite ($-z$ -panel), and the top panel ($+x$ - or $-x$ -panel depending on the definition of the body-fixed coordinate frame), while the panels where the solar arrays are mounted ($+y$ - and $-y$ -panel) as well as the bottom panel are not illuminated. The satellite body orientation with respect to the Sun can thus solely be parameterized by the Sun-elongation angle ε (the angle between Sun and center of the Earth as seen by the satellite). While the acceleration caused by radiation pressure exerted on the solar arrays remains constant and points away from the Sun for nominal attitude, the magnitude and direction of the acceleration caused by the satellite body vary with the Sun elongation angle or, equivalently, with argument of latitude u , that is, with the position angle of the satellite along its orbit.

As the average satellite cross-section exposed to the Sun as well as the amplitude of the required yawing motion depends on the elevation β of the Sun above or below the orbital plane, perturbations will repeat with the period the Sun takes from successive crossings through the same orbital plane in the northward direction. This so-called draconitic period is several days shorter than 1 year because the orbital planes perform a retrograde rotation due to gravitational perturbations caused by the Earth's oblateness. Neglecting the fact that the Sun is not moving along the celestial equator but along the ecliptic and using the precession rates from (3.26) we obtain the mean draconitic periods for different GNSS constellations displayed in Table 3.3. They correspond to the repeat periods of the Sun with respect to the respective GNSS orbit constellation. If we consider that the Sun is moving along the ecliptic, we realize that the draconitic period also depends on the

RAAN Ω of the orbital plane. The draconitic periods for individual orbital planes may thus differ from the values given in Table 3.3 by up to ± 30 days.

Two broad classes of GNSS radiation pressure models may be distinguished, physical models derived from satellite surface properties and empirical models derived from orbit analysis. Models of the first class typically provide the perturbing acceleration vector for given satellite orientation using mathematical functions, which approximate the accelerations derived from detailed physical space vehicle models. The first available models derived from physical characteristics of GPS satellites were the ROCK4 and ROCK42 models for the Block I and Block II/IIA vehicles developed by the spacecraft manufacturer Rockwell International and IBM. While the ROCK-S models consider the solar radiation, the ROCK-T models include in addition the thermal reradiation of the satellites. These models were approximated by Fourier series in the Sun elongation angle ε by [3.25] and called T10 and T20, respectively, [3.26] used the same approach to develop the T30 model for the Block IIR satellites based on a detailed spacecraft model from the manufacturer Martin Marietta while [3.27] developed an improved radiation pressure model for the same satellite type which includes satellite shadowing effects, Earth albedo radiation, thermal reradiation, and antenna radiation thrust.

Detailed physical models for GNSS satellites are developed at University College London. Ziebart and Dare [3.28] describe a radiation pressure model for the old GLONASS II v satellites based on ray tracing of a detailed model of the complexly shaped satellites. The model takes into account also shadowing effects and reflected radiation striking another part of the satellite [3.29]. More recently, similar models have been developed for Block IIR satellites [3.30, 31].

Empirical radiation pressure models include parameters that are adjusted in the orbit determination process. One of the most popular empirical models used within the International GNSS Service (IGS [3.32]; Chap. 33) is the empirical CODE (Center for Orbit Determination in Europe) orbit model (ECOM) developed

Table 3.3 Precession rate and mean draconitic period for the different GNSS constellations

System	Orbit precession ($^{\circ}$ /year)	Draconitic period (days)
GPS	− 15.16	351.4
GLONASS	− 12.10	353.4
BeiDou (MEO)	− 11.90	353.6
Galileo	− 9.69	355.7
QZSS	− 3.65	361.6
BeiDou (IGSO)	− 2.81	362.4
NavIC (IGSO)	− 4.36	360.9

at University of Bern [3.33]. The model represents the components of the solar radiation acceleration in a Sun-oriented reference frame with the unit vector of the first axis e_D pointing from the satellite to the Sun, the second axis e_Y pointing along the solar panel axis, and the third axis e_B completing a right-handed triad, or more specifically

$$\begin{aligned} e_D &= \frac{\mathbf{r}_\odot - \mathbf{r}}{\|\mathbf{r}_\odot - \mathbf{r}\|}, \\ e_Y &= \frac{e_D \times \mathbf{r}}{\|e_D \times \mathbf{r}\|}, \\ e_B &= e_D \times e_Y, \end{aligned} \quad (3.31)$$

where \mathbf{r} and \mathbf{r}_\odot represent the geocentric position vectors of the satellite and the Sun, respectively. This so-called DYB-reference frame thus rotates around the direction pointing to the Sun as the satellite performs its yaw-steering motion while orbiting the Earth. All three components of the radiation pressure acceleration vector with respect to the axes of this frame are represented by a constant term and a harmonic sine–cosine-term parameterized by the argument of latitude u of the satellite

$$\begin{aligned} D(u) &= D_0 + D_c \cos u + D_s \sin u, \\ Y(u) &= Y_0 + Y_c \cos u + Y_s \sin u, \\ B(u) &= B_0 + B_c \cos u + B_s \sin u. \end{aligned} \quad (3.32)$$

This first-order Fourier decomposition of the radiation pressure acceleration thus contains nine empirical parameters that are estimated in course of the orbit determination procedure. Due to the fact that once-per-revolution terms in the D and Y direction particularly strongly correlate with the orientation of the orbital plane [3.34], thus leading to an instability of the orbit orientation, in general only five of the nine parameters are estimated in the orbit determination process. This reduces the ECOM model to

$$\begin{aligned} D(u) &= D_0, \\ Y(u) &= Y_0, \\ B(u) &= B_0 + B_c \cos u + B_s \sin u. \end{aligned} \quad (3.33)$$

In addition to these five empirical parameters small and constrained velocity changes (stochastic pulses) are introduced for each GNSS satellite orbit at noon and at midnight in order to cope with residual orbit modeling deficiencies.

An a priori model that includes additional acceleration terms in the satellite body-fixed x - and z -directions

is described in [3.35]

$$\begin{aligned} D &= D_0, \\ Y &= Y_0, \\ B &= B_0, \\ Z(\Delta u) &= Z_1 \sin \Delta u, \\ X(\Delta u) &= X_1 \sin \Delta u + X_3 \sin 3\Delta u. \end{aligned} \quad (3.34)$$

Here $\Delta u = u - u_s$ represents the argument of latitude with respect to the Sun. This angle is related to the satellite position angle μ from the midnight point as defined in Fig. 3.19 through $\Delta u = \mu + \pi$. The six parameters of the model are in their turn parameterized as functions of the Sun elevation angle β above the orbital plane. The model thus includes a total of 18 parameters which were adjusted using several years of CODE final orbits. Similar models were later developed for the newer GPS block types as well as for GLONASS satellites. These models are used by several International GNSS Service (IGS) Analysis Centers as a priori models together with the five-parameter ECOM model (3.32).

Using a similar approach, empirical GPS solar pressure models (GSPM) were also developed at Jet Propulsion Laboratory (JPL). The GSPM.97 model by [3.36] was available for GPS Block IIA satellites. It was improved and extended to GPS Block IIR satellites by [3.37] based on more than 4 years of JPL orbits. The GSPM.04 model represents the acceleration components in the body-fixed frame as a truncated harmonic series with the Sun-elongation ε as argument

$$\begin{aligned} X(\varepsilon) &= X_1 \sin \varepsilon + X_2 \sin 2\varepsilon + X_3 \sin 3\varepsilon \\ &\quad + X_5 \sin 5\varepsilon + X_7 \sin 7\varepsilon, \\ Y(\varepsilon) &= Y_1 \cos \varepsilon + Y_2 \cos 2\varepsilon, \\ Z(\varepsilon) &= Z_1 \cos \varepsilon + Z_3 \cos 3\varepsilon + Z_5 \cos 5\varepsilon. \end{aligned} \quad (3.35)$$

Some of the 10 coefficients (Y_1 and X_2) are functions of the Sun-elevation angle β . Separate sets of model parameters are provided for GPS Blocks IIA and IIR. The model is used as a priori model while in addition a constant acceleration bias in the Y -direction and a constant scale parameter in the Sun direction as well as stochastic scale variations in the body-fixed coordinate axes are adjusted during orbit determination [3.38].

An alternative approach was developed by [3.39]. This adjustable box-wing model is based on a simple satellite box and solar panel array with defined geometrical dimensions. Optical properties of the illuminated surfaces are estimated during the orbit determination process, specifically the combined optical properties for the solar panels as well as the sum of absorption and diffusion coefficients and the reflection coefficients for

the $+x$ -, $+z$ - and $-z$ -surfaces of the satellite. To cope with correlations between the parameters the reflection coefficients are tightly constrained to a priori values. In addition to these parameters, an acceleration bias along the solar panel axis and a solar panel rotation lag angle are estimated during orbit determination. This model with a total of nine solve-for parameters can thus be considered as an empirical radiation pressure model with physical interpretation of the estimated parameters in terms of surface properties. Together with an improved satellite attitude model, the adjustable box-wing model results in a significant reduction of the impact of orbit modeling deficiencies at draconitic frequencies on geodetic time series such as station coordinates or apparent geocenter motion [3.40, 41].

Motivated by the finding that the classical empirical CODE orbit model (3.32) indicates deficiencies for GLONASS satellites [3.42], an enhanced version of the ECOM model was developed in [3.43].

$$\begin{aligned} D(u) &= D_0 + D_{2c} \cos 2\Delta u + D_{2s} \sin 2\Delta u \\ &\quad D_{4c} \cos 4\Delta u + D_{4s} \sin 4\Delta u, \\ Y(u) &= Y_0, \\ B(u) &= B_0 + B_c \cos \Delta u + B_s \sin \Delta u. \end{aligned} \quad (3.36)$$

The model includes additional empirical coefficients parameterizing higher frequency variation of the direct solar radiation acceleration in order to cope for the fact that the GLONASS satellites have an elongated shape causing more prominent variation of the satellite body's cross-section exposed to the Sun while orbiting the Earth. The argument Δu has the same interpretation as in (3.34).

For Galileo satellites orbit modeling deficiencies could also be identified when using the classical ECOM radiation pressure model (3.32). In fact, satellite laser ranging (SLR) residuals show prominent once-per-revolution variations with an amplitude depending on the Sun elevation angle β and reaching up to 20 cm [3.44]. Realizing that these variations originate from the notably rectangular shape of the Galileo in-orbit validation (IOV) satellites with a ratio of about 2 : 1 for the main body axes, an enhanced radiation pressure model was developed in [3.45]. This model complements the empirical CODE 5-parameter model (3.32) by a simple box-wing a priori model whose optical surface parameters were estimated based on Galileo tracking data spanning half a year.

Early Earth radiation models were developed by [3.20]. These models consider the perturbing impact of solar radiation reflected in the visible light on the surface of the Earth (albedo) and infrared radiation emitted by the Earth. The inclusion of these perturbations significantly reduces the orbit prediction errors for

GPS satellites [3.30]. Inclusion of Earth radiation and antenna thrust reduces the GPS orbit radius by 2 cm, resulting in a corresponding reduction of the observed bias of SLR residuals [3.21, 46].

Also, thermal imbalance in the satellite body and solar panels causes orbit perturbations through thermal re-radiation. Models for GPS satellites were developed by [3.47–49]. It was already shown by [3.47] that the impact of thermal re-radiation on orbit prediction reaches 10 m after 1 week for eclipsing GPS satellites. [3.48] demonstrates a reduction of the error for a 12 h orbit prediction of GPS Block IIR satellites from 2.7–3.0 m to 0.6 m in the alongtrack direction.

3.2.5 Long-Term Evolution

For operational reasons, the ground tracks of all GNSS constellations repeat after a time period, which is specific for each system (Table 3.1). While, for example, the GPS satellites perform two revolutions in one sidereal day, the Galileo satellites perform 17 revolutions in 10 d. For an Earth-fixed observer the GPS constellation thus repeats after one sidereal day and the Galileo constellation after 10 d. In general, the ground track of a satellite repeats, if its revolution period is commensurable with the rotation of the Earth. This means that the ratio of the revolution period U and the length of the sidereal day T , or, equivalently, the ratio of the Earth rotation rate ω_\oplus and the satellite's mean motion is an irreducible fraction

$$\frac{U}{T} = \frac{\omega_\oplus}{n} = \frac{K}{N} \quad (3.37)$$

of two integer numbers N and K . The satellite then performs N revolutions in K sidereal days. K is also called the orbit repeat cycle.

As a satellite on a ground track repeatable orbit appears again at the same location above the surface of the Earth after N revolutions, it encounters the same acceleration from the inhomogeneous gravity field of the Earth ever again. This periodic perturbation may result in a resonance effect. The most important resonant perturbation appears in the osculating semimajor axis, which results in a changed mean motion of the satellite. As a consequence, the satellite drifts away from its nominal position along the orbit.

These resonant perturbations are caused by specific terms in the Earth's gravity potential (3.26) with characteristic degree n and order m that depend on the integer numbers N and K of the repeat orbit [3.11]. For larger N potential terms of higher degree n are responsible for the perturbations. As the perturbing accelerations are decreasing with a power of $n+2$ of the satellite's distance, the resonant perturbations are much smaller for

satellites with larger number of revolutions N per cycle K .

For geostationary satellites, the resonance is particularly strong, since they are apparently fixed above a given location along the Earth's equator and thus experience a constant perturbation by the Earth's gravity field. Theory shows that the geopotential term $(n, m) = (2, 2)$, that is, the ellipticity of the Earth's equator, is responsible for the largest perturbation in this case. Indeed, GEO satellites require frequent station keeping maneuvers in order to keep them in the required longitude box of 0.1° , since the semimajor axis may drift up to 150 m/d, which causes a corresponding acceleration in longitude [3.50].

GEO satellites are, for example, used within the BeiDou constellation. To compensate the semimajor axis drift and the associated change in longitude, these satellites perform an orbit correction maneuver about once every three weeks. Figure 3.11 shows the geographic longitude for the satellite C01. Overlaid by shortperiodic perturbations, the satellite experiences a constant acceleration in the westward direction, which is periodically corrected by a maneuver changing the mean drift from westward to eastward. In a time interval of 200 d, a total of eight maneuvers can be observed. Also, IGSO satellites perform regular maneuvers, although only about two per year.

For satellites at a geostationary distance, the perturbing accelerations from Sun and Moon are of similar magnitude as the accelerations caused by the oblateness of the Earth. The combined perturbation results

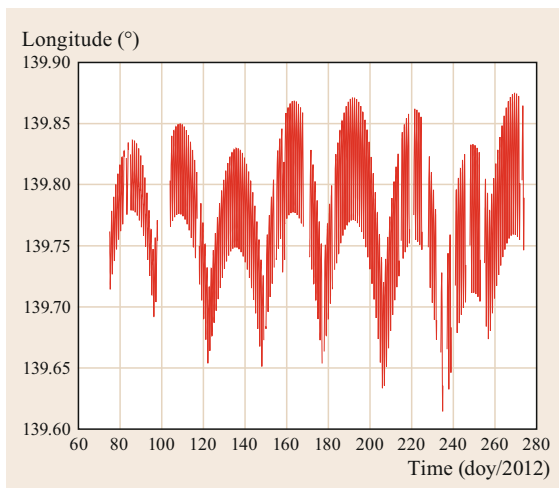


Fig. 3.11 Longitude of the geostationary BeiDou satellite C01 showing, apart from short-periodic variations, a resonant acceleration in the westward direction. The resulting mean longitude drift is periodically corrected by a station keeping maneuver

in a precession of the orbital planes around the so-called Laplace plane, a plane tilted around the line of equinoxes by about 7.5° from the equatorial toward the ecliptic plane. The consequence for geostationary satellites is an increase of the orbit inclination from 0° to 15° and back in a period of about 50 years [3.51]. To compensate the drift in inclination, geostationary satellites typically perform inclination maneuvers. For BeiDou GEO and IGSO satellites, however, no such maneuvers are observed. As a result, the BeiDou GEO satellite have a small and increasing inclination of a few degrees.

Orbits of GPS satellites are in a 2 : 1-commensurability with Earth rotation. Figure 3.12 shows the ground track of a selected GPS satellite together with the geopotential term with degree and order $(n, m) = (3, 2)$. This term causes the largest resonant perturbation as can be recognized from the perturbing accelerations exerted by this term, which repeat after every half revolution of the satellite. As an example, this geopotential term causes an acceleration in the positive alongtrack direction (red arrows) at the peak Northern and Southern latitudes. Drift rates in semimajor axis caused by this potential term may reach 6 m/d. The drift in semimajor axis caused by all resonant geopotential terms may reach 10 m/d for GPS satellites depending on the orbit eccentricity. Figure 3.13 shows the total resonant drift in semimajor axis for all GPS satellites together with the respective maximum possible value.

As a consequence, GPS satellites require regular station keeping maneuvers to cope with this resonant perturbation. Figure 3.14 displays the continuous growth of the mean semimajor axis for GPS satellite G04 due to resonant perturbations, which is interrupted by sudden decreases induced by maneuvers at about yearly intervals. GPS satellites in fact perform station-keeping maneuvers at an average rate of about 0.6 maneuvers per year (Fig. 3.15).

If no maneuvers were performed, the osculating elements would exhibit long-periodic variations with periods of 8 years or longer. The mean semimajor axis may show variations of up to 10 km while the deviation from the nominal slot position may reach $\pm 180^\circ$. Figure 3.16 shows the development of the mean semimajor axis and the deviation from the nominal slot position for PRN G04 from a numerical integration without applying any maneuver. In this example, the mean semimajor axis varies between about ± 4 km and the deviation of the nominal slot position reaches 210° . Not applying maneuvers would thus significantly affect the regular distribution of the GPS satellites in the constellation.

While GPS satellites require regular maneuvers, MEO satellites from other constellations do not need maneuvers as resonant perturbations can be neglected due to the much higher number N of revolutions per cy-

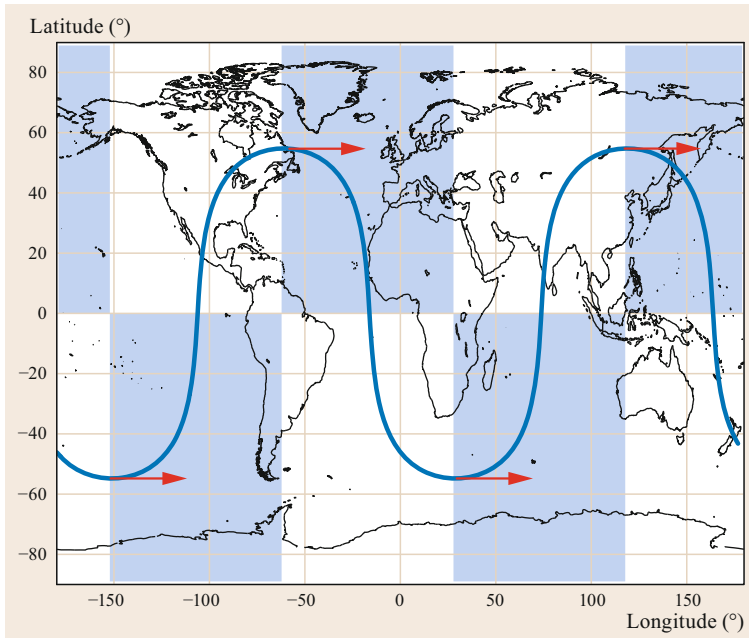


Fig. 3.12 Ground track of a GPS satellite together with the geopotential term with degree and order (3, 2). Arrows indicate the direction of acceleration at the orbit positions of minimum and maximum latitude

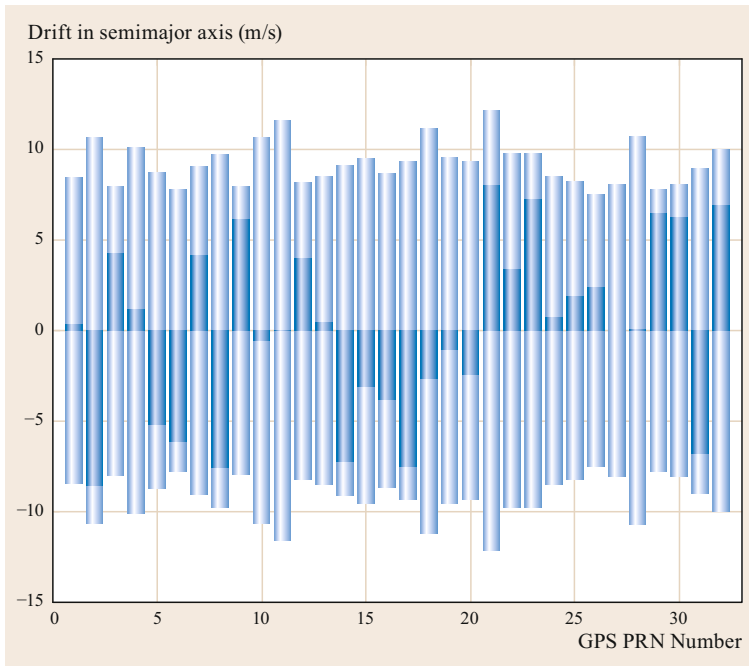


Fig. 3.13 Drift in semimajor axis for all GPS satellites (PRN numbers for July 2015). *Light*: maximum possible drift. *Dark*: actual total drift

cle for all of them. In fact, the gravity potential terms responsible for resonant perturbations have a minimum degree of 17 for GLONASS and Galileo and of 13 for BeiDou MEO satellites.

On the longer term, resonant perturbations of the eccentricity caused by accelerations from Sun and Moon and zonal terms of the geopotential have to be con-

sidered for MEO satellites as well [3.52–54]. These resonances occur when the secular motion of the lines of nodes and apsides becomes commensurable with the motion of Sun and Moon. Depending on the initial orbit orientation, these inclination-dependent resonances may lead to a quasi secular grow of the eccentricity, raising the orbit apogee and lowering the perigee, as

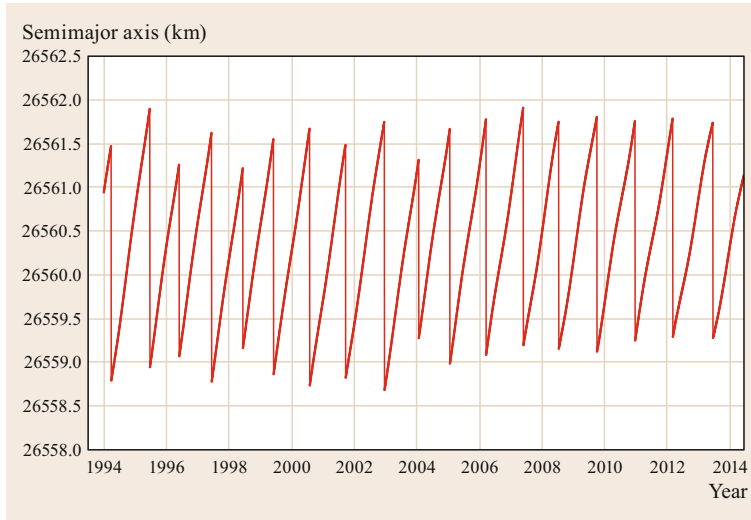


Fig. 3.14 Semimajor axis for GPS satellite PRN04 (space vehicle number SVN 34) showing a maneuver about once per year

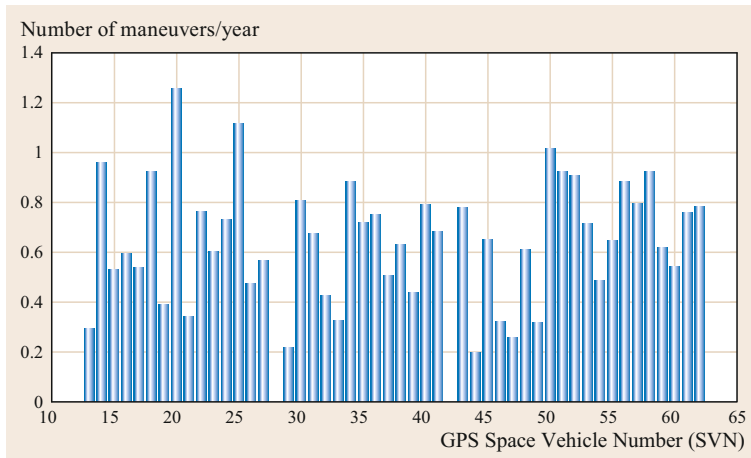


Fig. 3.15 Number of maneuvers per year performed by GPS satellites in the time period from January 1994 to July 2015. Repositioning maneuvers are not included

soon as, after end of life, orbit correction maneuvers are no longer performed.

Decommissioned GNSS satellites remain in orbit. They are reorbited into a *graveyard* orbit above or below the operational orbits in order to minimize the collision risk with operational satellites of the constellation. The perigee of the orbit of decommissioned GPS satellites is raised by about 1000 km, for Galileo satellites reorbiting policy consists of raising the orbital latitude by 300 km at end-of-life [3.56].

Due to the resonant growth of eccentricity reorbited satellites of a GNSS constellation may penetrate the operational altitude of a neighboring GNSS constellation within only a few decades [3.57, 58]. It is even discussed to exploit the resonant eccentricity build-up to lower the perigee of decommissioned GNSS satellites so far that they re-enter the atmosphere [3.58–60].

3.2.6 Orbit Accuracy

Precision and accuracy of orbits may be assessed with different methods each with its individual advantages and weaknesses [3.44, 61].

Precise orbit products are typically made available as daily files (Chaps. 33 and 34). Internal consistency of an orbit product may then be assessed by the analysis of the orbit discontinuities at day boundaries. This measure of internal quality depends on the arc length originally used for computing the orbits. Smaller day boundary discontinuities are, e.g., expected if the daily orbits are generated as the middle day of a three-day-arc.

Similarly, the precision of an orbit product can be assessed by adjusting a long orbit arc through orbit positions of successive one-day-arcs. The root mean square (RMS) of the orbit residuals then serves as

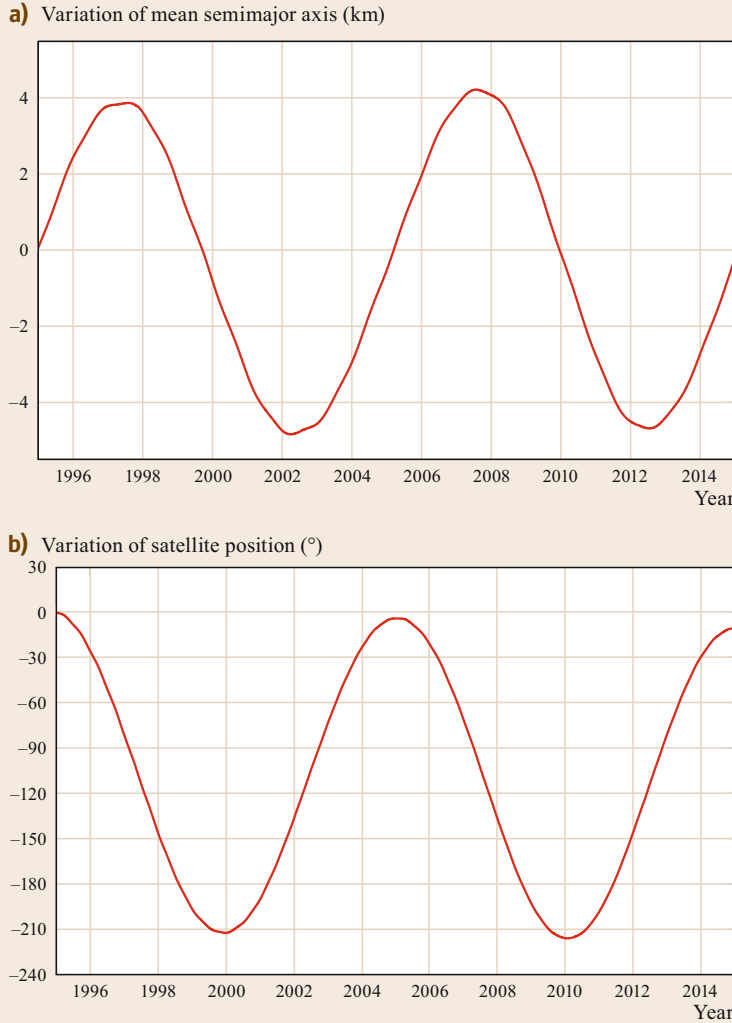


Fig. 3.16a,b Orbit of GPS satellite PRN04 (SVN 34) numerically integrated over 20 years without maneuver. **(a)** Mean semimajor axis. **(b)** Deviation of satellite position from nominal slot

a quality indicator. The indicator may however be biased toward the orbit model used as reference for the long-arc calculation. Also for this method optimistic values may be obtained for daily arcs extracted from longer arcs.

The comparison of orbits of the same satellites provided by different providers allows to assess the consistency of orbits computed using different software, analysis strategies, and subsets of the tracking network. The method allows the assessment of modeling differences between the providers while common biases remain undetected.

For satellites tracked by the SLR (satellite laser ranging) stations of the international laser ranging service [3.62]. Residuals of the ranges observed with this independent measurement technique versus. ranges computed from the orbits allow an assessment of the

orbit accuracy. Due to the high altitude of the GNSS satellites mainly the radial component of the orbit is validated. In addition, the GNSS satellites have to be equipped with laser retro-reflectors with known location with respect to the satellite's center of mass, and knowledge of the satellite attitude is required.

Only two (by now decommissioned), GPS Block IIA satellites, SVN 35 and 36, are equipped with laser retro-reflectors, while all satellites from all other GNSS constellations carry onboard retro-reflectors. New GPS satellites will again be equipped with retro-reflectors starting with GPS III SV-9 [3.63].

Orbits of GPS satellites provided by the IGS show a high degree of consistency and accuracy (Chaps. 33 and 34). The weighted RMS of the differences between the GPS orbits delivered by all IGS Analysis Centers and the final IGS orbit product is today be-

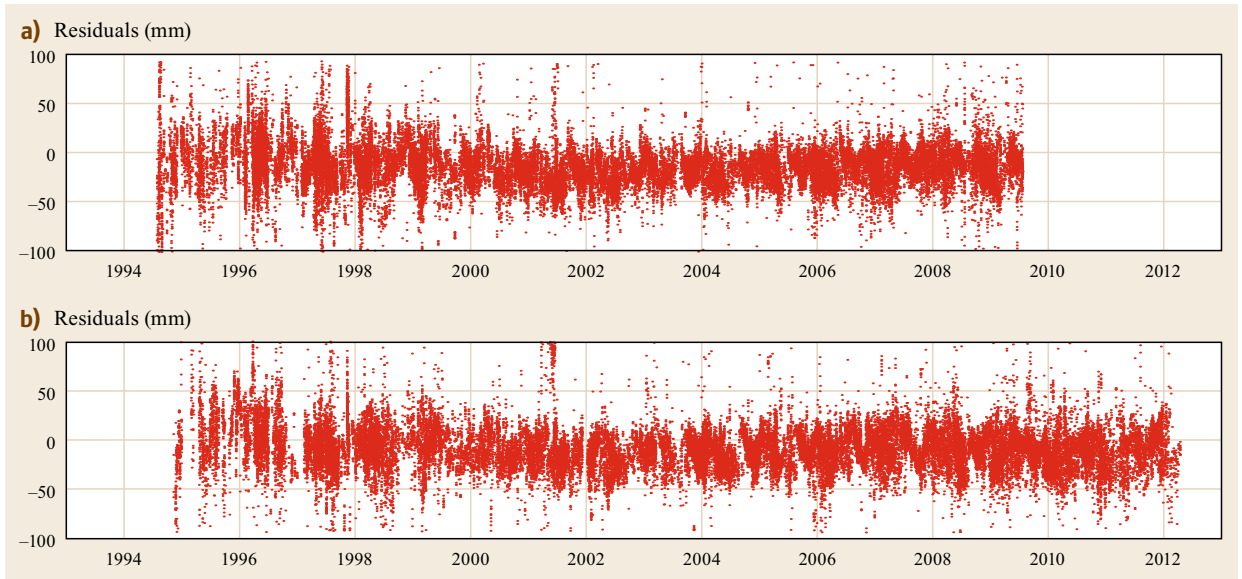


Fig. 3.17a,b Laser residuals based on reprocessed IGS orbits [3.55] for the two GPS satellites SVN 35 (a) and 36 (b) that are equipped with laser retro-reflectors

low 2 cm [3.61] despite the fact that the individual Analysis Center orbits are computed using different software packages, analysis strategies, and tracking station selections. This consistency corresponds with an average 1-day boundary orbit misfit for noneclipsing satellites of 21 mm [3.61]. Spectral analysis of daily orbit overlap time series of IGS orbits shows signals with amplitudes at the centimeter level with draconitic, fortnightly, and weekly periods that originate probably from orbit modeling issues and inaccuracies of the sub-daily Earth rotation model [3.64].

SLR residuals to the two GPS satellites equipped with laser retro-reflectors (Fig. 3.17) show a standard deviation of 19 mm (SVN 35) and 25 mm (SVN 36) [3.65]. New results by [3.66] based on reprocessed CODE orbits show an SLR standard deviation of 19 mm for the two satellites. Systematic SLR biases which were originally at a level of -5 to -6 cm (negative bias indicating that the SLR measurements show shorter ranges) [3.34] could be reduced to -3 cm [3.65] and finally to 13 mm [3.65] using improved orbit models and more accurate retro-reflector offset values.

Precise orbits of the Russian GLONASS satellites delivered by the IGS have a slightly reduced qual-

ity compared to the GPS final orbits. This is due to a reduced tracking network compared to the IGS GPS network, ambiguity resolution for satellite-specific frequencies, and orbit modeling issues. SLR RMS values for GLONASS-M satellites are between 30 and 40 mm with an average value of 35 mm and a vanishing bias [3.66].

For the new GNSSs, no orbits at a comparable quality are yet available. For the Galileo IOV satellites SLR residuals of up to 20 cm with a standard deviation of about 8–9 cm and a mean bias of about 4–5 cm were obtained originally [3.44]. With an improved solar radiation pressure model, the standard deviation and bias of the SLR residuals could be reduced to some 5 cm and -3 cm, respectively [3.45], Sect. 3.2.4. BeiDou MEO and IGSO orbits accuracies are currently at a level of about 2 dm while BeiDou GEO orbits are about 1 order of magnitude worse [3.67, 68]. In particular, the along-track component of the GEO orbits is difficult to determine because geostationary satellites perform only a small motion with respect to the ground stations. This causes strong correlations of the satellite longitude with pseudorange biases and phase ambiguities.

3.3 Broadcast Orbit Models

A key element of each satellite navigation system consists in the provision of broadcast orbit information, which enables the receiver to compute the position and

velocity of GNSS satellites in the constellation. Such information can be used to assess the visibility and tracking conditions when allocating the tracking chan-

nels and thus helps to speed up the initial acquisition. More importantly, however, accurate orbit information is required to compute modeled pseudoranges and pseudo-range rates as part of the navigation solution. In accord with these needs, two types of parameter sets are commonly distinguished:

- The *almanac* provides coarse orbit information with a typical accuracy at the 1 km level [3.69] to support the acquisition process. Each satellite transmits the orbit parameters of all satellites in the constellation along with auxiliary health and status information.
- The *ephemeris* provides (sub-)meter level orbit information for use in the position and velocity computation. Each satellite transmits only the ephemeris data for itself, thus allowing for a shorter total message length and a higher repeat rate.

Both almanac and ephemeris data are accompanied by satellite clock information (offset, drift, and, optionally, drift rate; Chap. 5), even though this information is solely required for the computation of the navigation solution. Detailed specifications are given in the interface control documents (ICDs) of the GPS [3.69], GLONASS [3.70], Galileo [3.71], BeiDou [3.72], and QZSS [3.73] systems, which provide a comprehensive description of the navigation message format and contents along with instructions for the use of all data.

Within this section, the almanac and ephemeris models employed in current GNSSs are discussed and basic algorithms are presented in a harmonized form. Despite large commonalities, care must be taken, though, that individual constellations employ different coordinate system realizations as well as different physical constants. Considering the latest reference frame realizations in use today, differences between constellations cause position offsets at the few centimeter level and can thus be neglected in comparison with the inherent precision of the broadcast ephemerides [3.74]. On the other hand, care should be taken to make proper use of the constellation-specific physical constants as defined within the respective ICDs (Table 3.4). This is particularly true for the Earth's gravitational coefficient, since even subtle GM_{\oplus} difference would affect the computed mean motion and create position errors that grow linearly over time.

Table 3.4 Physical parameters of GNSS almanac and ephemeris models

System	GM_{\oplus} (m ³ /s ²)	ω_{\oplus} (rad/s)
BeiDou	$398\,600.4418 \cdot 10^9$	$7.2921150 \cdot 10^{-5}$
Galileo	$398\,600.4418 \cdot 10^9$	$7.2921151467 \cdot 10^{-5}$
GLONASS	$398\,600.4418 \cdot 10^9$	$7.292115 \cdot 10^{-5}$
GPS	$398\,600.5 \cdot 10^9$	$7.2921151467 \cdot 10^{-5}$
QZSS	$398\,600.5 \cdot 10^9$	$7.2921151467 \cdot 10^{-5}$

3.3.1 Almanac Models

A Keplerian-style orbit model with six independent orbital elements is used to describe the constellation status in the navigation messages of GPS, Galileo, BeiDou, and QZSS. Aside from a secular orbital plane rotation caused by the Earth-oblateness, the corresponding almanac model does not consider orbital perturbations.

As summarized in Table 3.5, seven orbital parameters are provided to the user in addition to the almanac reference epoch t_a . These have originally been introduced in GPS, but were later inherited by most other GNSSs for improved communality. Instead of the semi-major axis a , its square-root is historically given to save the evaluation of a transcendental function in the computation of the mean motion

$$n_0 = \frac{\sqrt{GM_{\oplus}}}{\sqrt{a}^3} \quad (3.38)$$

and the mean anomaly

$$M = M_0 + n_0(t - t_a) \quad (3.39)$$

at the epoch of interest t . The inclination $i = i_{\text{ref}} + \delta_i$ is referred to a reference value i_{ref} (which varies from constellation to constellation) and only the difference is provided to optimize the number of data bits required for transmitting the almanac.

A specific feature of the GPS/GNSS almanac model is the direct computation of orbital positions in an Earth-fixed reference frame. Instead of the right ascension Ω of the ascending node, its longitude Ω_0 relative to the Greenwich meridian at the beginning t_w of the calendar week is given in the almanac data set. Both quantities are related by

$$\Omega_0 = \Omega(t_a) - \Theta(t_w), \quad (3.40)$$

Table 3.5 GNSS almanac parameters

Parameter	Description
\sqrt{a}	Square root of semimajor axis
e	Eccentricity
δ_i	Inclination offset from reference value $i_{\text{ref}} = 54^\circ$ (GPS), 63° (GLONASS), 56° (Galileo), 54° (BeiDou MEO/IGSO), 0° (BeiDou GEO), 45° (QZSS)
Ω_0	Longitude of the ascending node at the weekly epoch
$\dot{\Omega}$	Rate-of-change of the right ascension of the ascending node
ω	Argument of perigee
M_0	Mean anomaly at reference epoch

where Θ denotes the Greenwich mean sidereal time. As a consequence of the Earth's oblateness, the inertial RAAN is not a constant, but exhibits a mean drift $\dot{\Omega}$ in the retrograde direction, which varies with inclination and orbital radius. It ranges from a minimum of roughly $-0.01^\circ/\text{d}$ for high-altitude, geosynchronous orbits to a maximum of about $-0.04^\circ/\text{d}$ for the GPS MEO constellation (Table 3.3).

Making use of the nodal rate provided in the almanac parameters, the Earth-fixed longitude λ_Ω of the ascending node at epoch t can be expressed as

$$\begin{aligned}\lambda_\Omega(t) &= \Omega(t) - \Theta(t) \\ &\approx \Omega(t_a) + \dot{\Omega}(t - t_a) \\ &\quad - \Theta(t_w) - \omega_\oplus(t - t_w) \\ &= \Omega_0 + \dot{\Omega}(t - t_a) - \omega_\oplus(t - t_w). \quad (3.41)\end{aligned}$$

Following the solution of Kepler's equation (3.11) for the eccentric anomaly E and the computation of the perifocal coordinates (x_p, y_p) from (3.9), the position of the GNSS satellite in the ITRF (or the respective constellation-specific, Earth-fixed frame) can thus be computed from the relation

$$\mathbf{r}_{\text{ITRF}} = \mathbf{R}_3(-\lambda_\Omega)\mathbf{R}_1(-i)\mathbf{R}_3(-\omega) \cdot \begin{pmatrix} x_p \\ y_p \\ 0 \end{pmatrix}. \quad (3.42)$$

Differentiation with respect to time t yields the associated expression

$$\begin{aligned}\dot{\mathbf{r}}_{\text{ICRF}} &= \mathbf{R}_3(-\lambda_\Omega)\mathbf{R}_1(-i)\mathbf{R}_3(-\omega) \begin{pmatrix} \dot{x}_p \\ \dot{y}_p \\ 0 \end{pmatrix} \\ &\quad - \begin{pmatrix} 0 \\ 0 \\ \omega_\oplus - \dot{\Omega} \end{pmatrix} \times \mathbf{r}_{\text{ICRF}}. \quad (3.43)\end{aligned}$$

for the velocity relative to the rotating ITRF, where (\dot{x}_p, \dot{y}_p) is the perifocal velocity obtained in (3.13).

For completeness, we note that the generic formulation of the almanac model given above is mathematically equivalent to the algorithms specified in the GPS, Galileo, BeiDou, and QZSS ICDs, but employs a slightly different notation and arrangement of the individual equations. Among others, the ICD formulation refers all times to the start-of-week, makes use of the true anomaly and avoids explicit matrix rotations. Furthermore, the ICDs lack instructions for the computation of the spacecraft velocity.

Other than the aforementioned constellations, the GLONASS system specifies an analytical almanac model, which considers both secular and short periodic perturbations to compute position and velocity

from a set of six mean orbital elements. While potentially more accurate than the GPS almanac model, the GLONASS model is substantially more complex and cannot be presented here in adequate detail. Interested readers are referred to the GLONASS ICD [3.70] for a comprehensive formulation, while a truncated version with adequate accuracy for practical purposes is described in [3.75].

3.3.2 Keplerian Ephemeris Models

A perturbed Keplerian orbit representation is used for the broadcast ephemeris model of the GPS, Galileo, BeiDou, and QZSS constellations. Other than GLONASS that builds on a numerical orbit propagation, the spacecraft position and velocity can directly be computed from the given orbital parameters at arbitrary epochs in the overall validity interval.

The most common ephemeris model has been established for use with the GPS LNAV and later inherited by most other constellations. It represents an extended version of the almanac model discussed earlier and introduces additional parameters for a refined orbit representation (Table 3.6). These account for differences in the mean motion relative to its Keplerian value,

Table 3.6 Parameters of Keplerian broadcast ephemeris models

Parameter	Description
$\sqrt{a}, \Delta a$	Square root of semimajor axis legacy navigation message (LNAV) or semimajor axis offset from reference value $a_{\text{ref}} = 26\,559\,710\text{ m}$ (GPS civil navigation message (CNAV)), $42\,164\,200\text{ m}$ (QZSS CNAV)
\dot{a}	Rate of change of the semimajor axis (CNAV)
Δn	Correction to mean motion
\dot{n}_0	Rate of change of mean motion (CNAV)
e	Eccentricity
i_0	Inclination at reference epoch
di/dt	Rate-of-change of inclination
Ω_0	Longitude of the ascending node at the weekly epoch
$\dot{\Omega}$	Rate-of-change of the right ascension of the ascending node (LNAV)
$\Delta\dot{\Omega}$	Rate-of-change of the right ascension of the ascending node relative to a reference value $\dot{\Omega}_{\text{ref}} = 4.68 \cdot 10^{-7} \text{ }^\circ/\text{s}$ (GPS/QZSS CNAV)
ω	Argument of perigee
M_0	Mean anomaly at reference epoch
$C_{\text{rc}}, C_{\text{rs}}$	Amplitude of (co)sine harmonic correction term to the orbital radius
$C_{\text{uc}}, C_{\text{us}}$	Amplitude of (co)sine harmonic correction term to the argument of latitude
$C_{\text{ic}}, C_{\text{is}}$	Amplitude of (co)sine harmonic correction term to the inclination

a drift of the orbital inclination and for perturbations in the radial, along-track, and cross-track direction with a twice-per-rev characteristics. To achieve even higher accuracy, the drifts of the semimajor axis and the mean motion are, furthermore, considered in the new CNAV transmitted with the GPS L2C and L5 signals [3.69, 76]. Also, the CNAV message provides the offset of the nodal drift from a reference value rather than the full value to achieve higher precision with a smaller number of data bits. All orbital elements and parameters refer within a broadcast navigation message refer to a common epoch t_e , but different epochs (and elements) apply for the LNAV and CNAV message types.

The ephemeris algorithm starts with computing the mean anomaly

$$M = M_0 + n(t - t_e) \quad (3.44)$$

at the epoch of interest from the semimajor axis

$$a = \begin{cases} (\sqrt{a})^2 & \text{(LNAV)} , \\ a_{\text{ref}} + \Delta a & \text{(CNAV)} , \end{cases} \quad (3.45)$$

and (perturbed) mean motion

$$n = \sqrt{\frac{GM_{\oplus}}{a^3}} + \begin{cases} \Delta n & \text{(LNAV)} , \\ \Delta n + \Delta \dot{n}(t - t_e) & \text{(CNAV)} . \end{cases} \quad (3.46)$$

Following the solution of Kepler's equation (3.11), the true anomaly

$$v = 2 \tan^{-1} \left(\sqrt{\frac{1+e}{1-e}} \tan \frac{E}{2} \right) \quad (3.47)$$

and the unperturbed argument of latitude $\bar{u} = \omega + v$ are obtained. The latter serves to evaluate the periodic corrections

$$\begin{aligned} \delta r &= C_{rs} \sin(2\bar{u}) + C_{rc} \cos(2\bar{u}) , \\ \delta u &= C_{us} \sin(2\bar{u}) + C_{uc} \cos(2\bar{u}) , \\ \delta i &= C_{is} \sin(2\bar{u}) + C_{ic} \cos(2\bar{u}) , \end{aligned} \quad (3.48)$$

and yields the perturbed values

$$\begin{aligned} r &= a(1 - e \cos E) + \delta r , \\ u &= \bar{u} + \delta u , \\ i &= i_0 + \frac{di}{dt}(t - t_e) + \delta i , \end{aligned} \quad (3.49)$$

of the radius, the argument of latitude, and the inclination. As in the almanac model, the Greenwich longitude of the ascending node is given by

$$\lambda_{\Omega} = \Omega_0 + \dot{\Omega}(t - t_e) - \omega_{\oplus}(t - t_w) , \quad (3.50)$$

where the nodal rate $\dot{\Omega}$ is either given directly in the ephemeris data (LNAV) or obtained from the reference value (CNAV)

$$\dot{\Omega} = \dot{\Omega}_{\text{ref}} + \Delta \dot{\Omega} . \quad (3.51)$$

Collecting the above results, the Earth-fixed position is finally given by

$$\mathbf{r}_{\text{ITRF}} = \mathbf{R}_3(-\lambda_{\Omega}) \mathbf{R}_1(-i) \begin{pmatrix} r \cos u \\ r \sin u \\ 0 \end{pmatrix} . \quad (3.52)$$

Complementary to the GNSS satellite position, its velocity is required as part of the navigation solution process, when determining the user velocity from observed range rate or Doppler values (Chap. 21). While undocumented in the ICDs, the velocity can again be found by differentiation of the above expressions for the position with respect to time t . The resulting expression are substantially more complex, though, and cannot be reproduced here in view of limited space. Interested readers are therefore referred to [3.77, 78], and [3.79], which provide consistent sets of equations for computing position, velocity, and, optionally, acceleration in the Earth-fixed reference frame based on the standard LNAV ephemeris model.

The basic GPS LNAV ephemeris model has likewise been adopted by Galileo and BeiDou. System-specific adaptations are limited to the use of individual time and reference systems and a slightly varying bit-field representation of the individual quantities [3.71, 72]. As a major addition, a special formulation has, however, been introduced for the specific case of geostationary BeiDou satellites in view of their moderate inclination with respect to the equator [3.72, 80]. The inclination i provided in the broadcast ephemeris for GEO satellites is therefore referred to an auxiliary plane with a 5° tilt relative to the Earth equator. Also, a different convention and origin apply for the specification of the longitude of the ascending node Ω . These result in the alternative expression

$$\begin{aligned} \mathbf{r} &= \mathbf{R}_3(\omega_{\oplus}(t - t_w)) \mathbf{R}_1(-5^\circ) \\ &\quad \times \mathbf{R}_3(-\Omega) \mathbf{R}_1(-i) \begin{pmatrix} r \cos u \\ r \sin u \\ 0 \end{pmatrix} \end{aligned} \quad (3.53)$$

with

$$\Omega(t) = \Omega_0 + \dot{\Omega}(t - t_e) - \omega_{\oplus}(t_e - t_w) . \quad (3.54)$$

Making use of the auxiliary plane, potential singularities can readily be avoided. Since the orbital plane of

the BeiDou GEO satellites is typically controlled to fall within less than about 2° from the equator, the inclination with respect to the auxiliary plane can never attain a zero value. It is emphasized, though, that the modified model is exclusively used for the geostationary satellites of the BeiDou constellation, whereas the standard model and parameterization are employed for the BeiDou MEO and IGSO satellites.

3.3.3 Cartesian Ephemeris Model

An alternative for distributing GNSS orbit information has been introduced by the Russian GLONASS system. Here, users are provided with a Cartesian state vector (i. e., position \mathbf{r} and velocity \mathbf{v}) at the ephemeris reference epoch t_e . The trajectory in the vicinity of this epoch can then be obtained by numerical integration of first-order equation of motion

$$\frac{d}{dt} \begin{pmatrix} \mathbf{r} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{v} \\ \mathbf{a} \end{pmatrix}. \quad (3.55)$$

To avoid the need for explicit reference system transformations, the equation of motion is expressed in the rotating, Earth-fixed reference frame. Accordingly, centrifugal and Coriolis terms are considered in the modeled acceleration

$$\begin{aligned} \mathbf{a} = & -GM_{\oplus} \frac{\mathbf{r}}{r^3} \\ & - \frac{3}{2} J_2 GM_{\oplus} \frac{R_{\oplus}^2}{r^5} \begin{pmatrix} \frac{x-5xz^2}{r^2} \\ \frac{y-5yz^2}{r^2} \\ \frac{3z-5z^3}{r^2} \end{pmatrix} \\ & + \omega_{\oplus}^2 \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} + 2\omega_{\oplus} \begin{pmatrix} +\dot{y} \\ -\dot{x} \\ 0 \end{pmatrix} \\ & + a_{\odot\oplus} \end{aligned} \quad (3.56)$$

next to the central component of the Earth gravity field and the Earth oblateness [3.70, 81]. Within this expression, the spacecraft position $\mathbf{r} = (x, y, z)^\top$ and the acceleration \mathbf{a} are consistently referred to an Earth-fixed reference frame aligned with the instantaneous equator and rotation axis. Third-body perturbations by the Sun and Moon are taken into account by the acceleration $a_{\odot\oplus}$, which is considered as constant over the ephemeris propagation interval (of typically ± 15 min). To simplify the user algorithm, $a_{\odot\oplus}$ is provided as part of the navigation message next to position and velocity at the reference epoch.

Based on the equation of motion and the given initial conditions, the position and velocity can be obtained at any time within the validity interval through

Table 3.7 Typical update interval of broadcast ephemerides for individual GNSSs and navigation message types

System	Type	Interval
GPS	LNAV	2 h
	CNAV	3 h
GLONASS		30 min
BeiDou		1 h
Galileo	INAV, FNAV	10–180 min
QZSS	LNAV, CNAV	15 min

numerical integration. While no particular integration method is specified within the GLONASS ICD, a fourth-order Runge–Kutta method is commonly recommended. It provides an approximation

$$\mathbf{y}(t+h) \approx \mathbf{y}(t) + \frac{h}{6} (\mathbf{k}_1 + 2\mathbf{y}_2 + 2\mathbf{y}_3 + \mathbf{y}_4) \quad (3.57)$$

with

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}(t, \mathbf{y}(t)), \\ \mathbf{k}_2 &= \mathbf{f}\left(t + \frac{h}{2}, \mathbf{y}(t) + \frac{h\mathbf{k}_1}{2}\right), \\ \mathbf{k}_3 &= \mathbf{f}\left(t + \frac{h}{2}, \mathbf{y}(t) + \frac{h\mathbf{k}_2}{2}\right), \\ \mathbf{k}_4 &= \mathbf{f}(t+h, \mathbf{y}(t) + h\mathbf{k}_3). \end{aligned} \quad (3.58)$$

A simplified form of a Cartesian trajectory model has, furthermore, been adopted for the various SBAS systems. Here, a set of position, velocity and acceleration values is provided at the given reference epoch, which describes a Taylor series approximation

$$\mathbf{r}(t) = \mathbf{r}(t_e) + \mathbf{v}(t_e)(t-t_e) + \frac{1}{2}\mathbf{a}(t_e)(t-t_e)^2 \quad (3.59)$$

of the spacecraft position over short time intervals [3.82, 83]. As in the GLONASS model, all quantities refer directly to the Earth-fixed coordinate system. The simplicity of the SBAS ephemeris model in comparison to the more elaborate GLONASS model is mainly achieved through frequent ephemeris updates as well as a continuous upload capabilities. Compared to the 24 h period of the geostationary SBAS satellites, the common update interval of about 4 min covers less than 1/300th of an orbit. Accordingly, higher-order terms can be neglected and a second-order expansion is sufficient to meet the desired accuracy.

3.3.4 Broadcast Ephemeris Generation and Performance

Consecutive sets of ephemeris parameters with update intervals of 10 min to 3 h (Table 3.7) are generated from a common predicted trajectory and uploaded as a com-

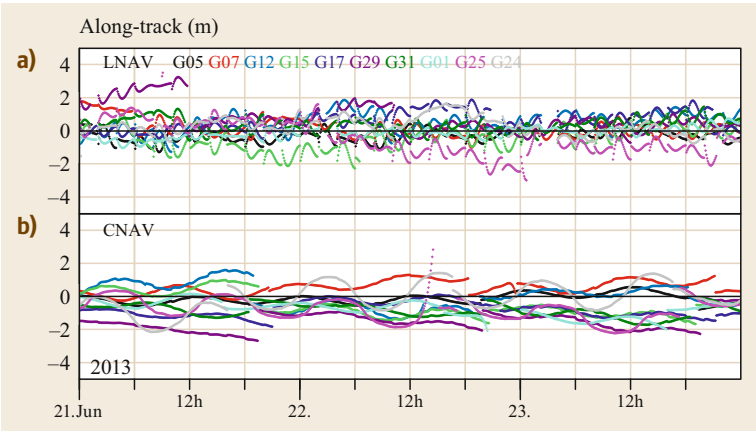


Fig. 3.18a,b Along-track position errors of GPS LNAV (a) and CNAV (b) navigation messages during the first CNAV transmission in June 2013 (after [3.74])

mon batch to the GPS satellite during a ground contact. As a minimum, uploads are performed once per day during nominal operations in GPS [3.84]. Even though consecutive ephemeris sets refer to the same trajectory between such uploads, small discontinuities will still be encountered when transitioning from one set to another due to the inherent approximation error. For the GPS LNAV message, such errors may amount to 0.5 m but are reduced to centimeter level in the new CNAV message [3.84, 85]. The smooth nature of the CNAV ephemeris can also be recognized from the comparison of LNAV and CNAV ephemeris errors shown in Fig. 3.18. While both ephemeris types exhibited similar overall errors during the early CNAV test phase, no discontinuities can be recognized at the (three hourly) updates in between the daily uploads.

Discontinuities at the 1 m level are also present in the GLONASS ephemeris as a result of the limited resolution (1 mm/s) of the velocity field within the GLONASS navigation message. The resulting approximation errors are still smaller than the overall ephemeris error budget but may in fact limit future performance improvements.

The quality of broadcast ephemeris data of GPS and GLONASS has been assessed in various studies [3.86–88], which demonstrated a continuous improvement over time. The performance figures for the two legacy systems are compared in Table 3.8 with early results for the new constellations BeiDou, Galileo

Table 3.8 RMS position errors radial (R), along track (T) and the cross-track (N) direction of broadcast ephemerides in 2013/14 (after [3.74])

System	Type	R (m)	T (m)	N (m)
GPS	LNAV	0.18	1.05	0.44
GLONASS		0.35	2.41	1.33
BeiDou (MEO+IGSO)		0.50	2.42	1.31
Galileo	INAV	0.63	2.65	2.29
QZSS	LNAV	0.48	1.42	0.92

and QZSS based on a comparison of broadcast and precise ephemerides over a 1 year interval. While the along-track component is least well determined in all systems, it only contributes an average of 9–15% to the user range error (URE) depending on the orbit altitude. Radial errors, which directly map into the URE are typically confined to less than 0.5 m. For completeness, it is emphasized that the values given Table 3.8 do not include the contribution of clock offset uncertainties in the broadcast ephemerides. Depending on the clock type (cesium, rubidium, or hydrogen-maser; Chap. 5) and quality, the errors of the broadcast clock will typically exceed the orbit uncertainty and therefore dominate the overall error budget. The total signal-in-space range error (SISRE), which measures the contribution of broadcast orbit and clock errors to the modeled pseudorange, presently amounts to 0.7 m for GPS but up to 2 m for other constellations [3.74].

3.4 Attitude

The discussion of orbital dynamics in the previous sections aims at describing the motion of the spacecraft center-of-mass (CoM) under the action of various external forces. However, knowledge of the CoM location alone is not sufficient for a precise modeling of GNSS observations. All radio signals transmitted by a GNSS satellite originate from the phase center of the antenna, rather than the CoM. The offset of the antenna from the CoM is a constant vector in a body-fixed coordinate system, but varies in space depending on the instantaneous orientation, or *attitude*, of the satellite body. Aside from the antenna offset, the changing orientation of the spacecraft relative to the observer may also affect the observed carrier phase of the transmitted signal through the so-called *phase wind-up* effect. A comprehensive discussion of both aspects is provided as part of the observation modeling in Chap. 19. Last, but not the least, solar radiation pressure modeling (Sect. 3.2.4) relies on a proper understanding of the orientation of the spacecraft body and the solar panels relative with respect to the orbit and the Sun.

The nominal orientation of a GNSS satellite is driven by a small set of requirements, which is largely independent of the particular system or satellite manufacturer. First, the boresight of the antenna must always be directed to the center of the Earth to maintain an optimum coverage and proper strength of the navigation signals. Secondly, the solar panels shall be aligned perpendicular to the Sun direction to maximize the pro-

jected area and thus the received solar energy. The solar panel rotation axis must therefore be oriented perpendicular to the plane spanned by the Sun and Earth direction. Finally, one of the satellite faces perpendicular to the antenna boresight and solar panel rotation axis should permanently point into the hemisphere opposite the Sun to facilitate thermal stabilization of the atomic clocks (mounted close to this *cool* panel).

The resulting attitude control mode is commonly known as *yaw-steering mode* and jointly applied by GPS, GLONASS, Galileo, BeiDou, and QZSS satellites in MEO or IGSO orbits. It is illustrated in Fig. 3.19, where the principal spacecraft axes have been labeled in accord with established conventions [3.89] of the international GNSS service (IGS):

- The $+x$ -, y -, and z -axis form a right-handed coordinate system attached to the satellite body.
- The $+z$ -axis coincides with the antenna boresight direction.
- The y -axis is parallel to the rotation axis of the solar panels. Furthermore, the $+y$ -direction is assigned such that the $+x$ -panel is illuminated by the Sun during nominal yaw-steering, while the $-x$ -panel is oriented toward deep space.

Maintaining the ideal GNSS attitude requires a permanent rotation of the spacecraft body about the Earth-pointing $+z$ -axis as well as a rotation of the solar panels about the $+y$ -axis.

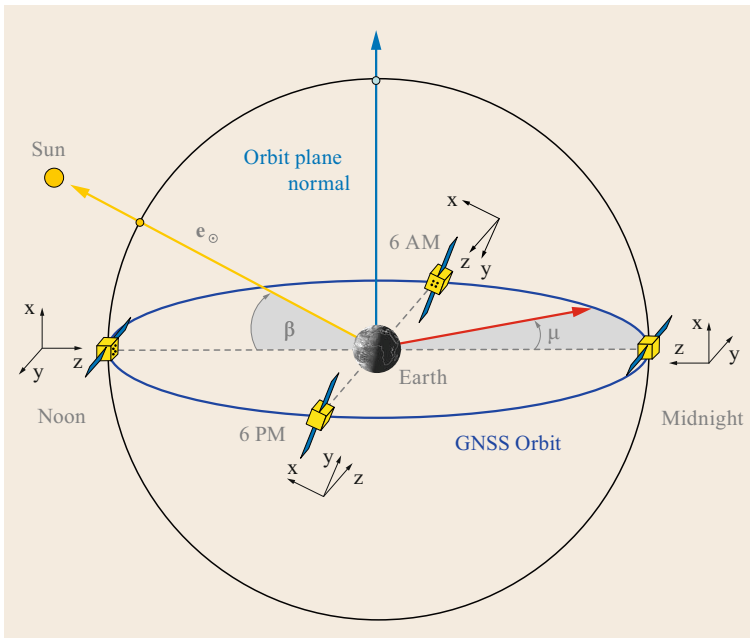


Fig. 3.19 GNSS satellite orientation in yaw-steering mode (after [3.89])

The required orientation of the spacecraft body is most easily described in an orbital reference frame aligned with the unit vectors

$$\begin{aligned} \mathbf{e}_R &= \frac{\mathbf{r}}{||\mathbf{r}||}, \\ \mathbf{e}_T &= \mathbf{e}_N \times \mathbf{e}_R, \\ \mathbf{e}_N &= \frac{\mathbf{r} \times \mathbf{v}}{||\mathbf{r} \times \mathbf{v}||}, \end{aligned} \quad (3.60)$$

in the radial, transverse, and normal direction (Fig. 3.20), which are defined by the instantaneous position \mathbf{r} and velocity \mathbf{v} of the GNSS satellite. The yaw-angle Ψ specifies the angle between the \mathbf{e}_T - and \mathbf{e}_x -axes for a right-handed rotation around the $+z/-R$ -axis. For $\Psi = 0^\circ$, the spacecraft $+x$ -axis is aligned with the transverse direction and $+y$ is oriented antiparallel

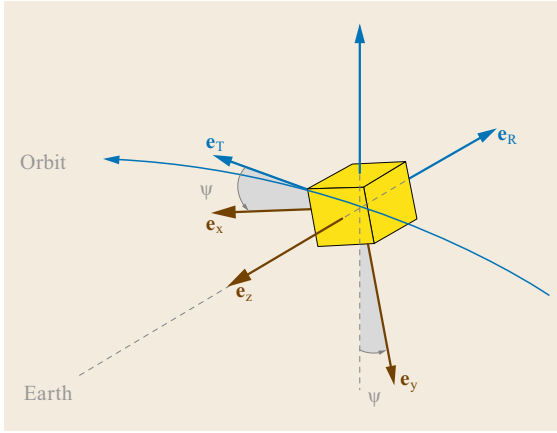


Fig. 3.20 Definition of the yaw-angle (after [3.89])

to the orbital angular momentum. Following [3.36], the nominal yaw angle in yaw-steering mode can be expressed as

$$\Psi = \text{atn2}(-\tan \beta, \sin \mu), \quad (3.61)$$

where β denotes the elevation of the Sun above the orbit plane and μ measures the orbit angle relative to the midnight point (Fig. 3.19). The nominal attitude of a GNSS satellite is thus fully determined by its orbital position and the direction of the Sun.

According to its definition, the yaw-angle is negative for positive Sun elevations and vice versa. Its variation with orbit angle is illustrated in Fig. 3.21 for various values of the Sun elevations. While Ψ remains close to $\pm 90^\circ$ for high β -angles, it varies between roughly 0° and $\pm 180^\circ$ whenever the Sun is close to the orbit plane. During these periods, which coincide with the eclipse season, the GNSS satellites need to perform rapid yaw-slews near orbit angles of $\mu = 0^\circ$ and $\mu = 180^\circ$. Depending on the design of the attitude control system (ACS), the actual yaw rate may be limited thus inhibiting a perfect yaw-steering during these *noon-* and *midnight-turns*. Furthermore, the lack of Sun-visibility during eclipses may result in nonnominal yaw angles during limited periods of time. Models describing the yaw-angle variation for the GPS Block IIA/IIR/IIF satellites as well as GLONASS-M satellites in low- β regimes and during eclipse transits have been developed by various authors ([3.36, 90–92] and [3.93], respectively) and are further discussed in Chap. 19.

To avoid the need for rapid yaw slews, the orbit-normal (or yaw-fixed) mode is employed by QZSS [3.94] and the Beidou MEO/IGSO satellites [3.95] as an alternative to yaw-steering during periods of low β -angles.

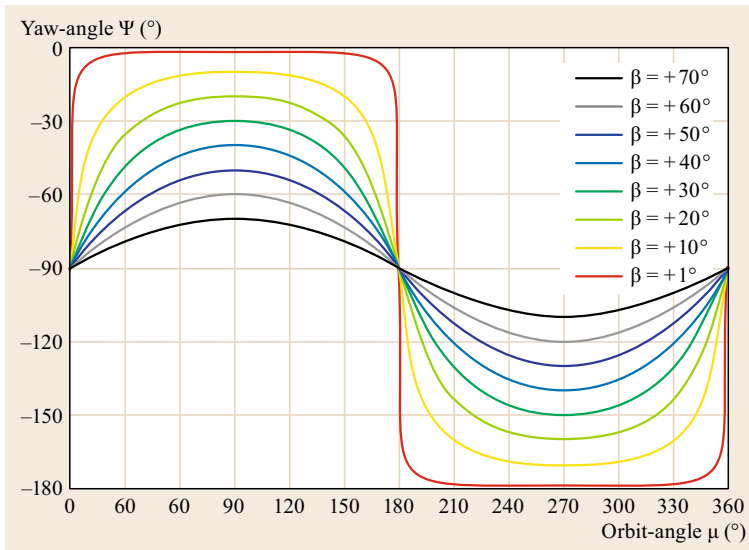


Fig. 3.21 Yaw-angle variation for positive β -angles

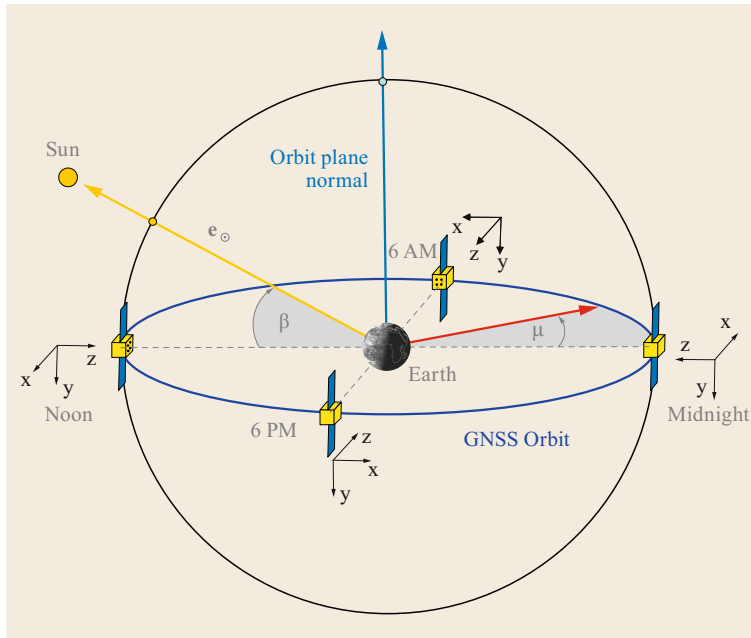


Fig. 3.22 GNSS satellite orientation in orbit-normal mode with a fixed yaw-angle $\Psi = 0^\circ$ (after [3.89])

In this mode, either the $+x$ -axis ($\Psi = 0^\circ$) or the $-x$ -axis ($\Psi = 180^\circ$) is permanently aligned with the transverse direction of the orbital frame (Fig. 3.22) and the solar panels rotation axis is kept perpendicular to the orbital plane. Accordingly, the effective cross-section of the solar panels is slightly decreased, which results in a small, but tolerable, reduction of the received power.

In the orbit-normal mode, the $-x$ -panel is no longer kept away from the Sun, but illuminated in turn with the $+z$ -, $+x$ -, and $-z$ -panels. Likewise, either the $+y$ - or $-y$ -

panel is permanently illuminated. Both effects need to be properly considered in the radiation pressure modeling and may constitute a challenge for precise orbit determination [3.96].

The transition from yaw-steering to orbit-normal mode takes place when $||\beta||$ drops below a limiting value of about 20° for QZS-1 [3.94] and about 4° for BeiDou. However, the exact instant of the mode switch is determined by the control center and may deviate slightly from the idealized values [3.97–99].

References

- 3.1 G. Beutler: *Methods of Celestial Mechanics* (Springer, Berlin 2005)
- 3.2 R. Fitzpatrick: *An Introduction to Celestial Mechanics* (Cambridge Univ. Press, Cambridge 2012)
- 3.3 W. Gellert, S. Gottwald, M. Hellwich, H. Kästner, H. Küstner: *The VNR Concise Encyclopedia of Mathematics*, 2nd edn. (Van Nostrand Reinhold, New York 1989)
- 3.4 D.A. Vallado: *Fundamentals of Astrodynamics and Applications*, 2nd edn. (Kluwer Academic, Dordrecht 2001)
- 3.5 O. Montenbruck, E. Gill: *Satellite Orbits – Models, Methods and Applications* (Springer, Berlin 2000)
- 3.6 P.J. Mohr, B.N. Taylor, D.B. Newell: CODATA recommended values of the fundamental physical constants: 2010, *J. Phys. Chem. Ref. Data* **41**(4), 043109 (2012)
- 3.7 G. Petit, B. Luzum: *IERS Conventions* (Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt 2010), IERS Technical Note No. 36
- 3.8 R.H. Battin: *An Introduction to the Mathematics and Methods of Astrodynamics* (AIAA, New York 1999)
- 3.9 D. Brouwer: Solution of the problem of artificial satellite theory without drag, *Astron. J.* **64**, 378–396 (1959)
- 3.10 I. Kozai: Second-order analytical solution of artificial satellite theory without air drag, *Astron. J.* **67**(7), 446–461 (1962)
- 3.11 W.M. Kaula: *Theory of Satellite Geodesy* (Blaisdell, Waltham 1966)
- 3.12 L.E. Cunningham: On the computation of the spherical harmonic terms needed during the numerical integration of the orbital motion of an artificial satellite, *Celest. Mech.* **2**(2), 207–216 (1970)
- 3.13 R. Pail, S. Bruinsma, F. Migliaccio, C. Förste, H. Goiginger, W.-D. Schuh, E. Höck, M. Reguzzoni, J.M. Brockmann, O. Abrikosov, M. Veicherts, T. Fecher, R. Mayrhofer, I. Krasbutter, F. Sansò, C.C. Tscherning:

- First GOCE gravity field models derived by three different approaches, *J. Geod.* **85**(11), 819–843 (2011)
- 3.14 N.K. Pavlis, S.A. Holmes, S.C. Kenyon, J.K. Factor: The development and evaluation of the Earth gravitational model 2008 (EGM2008), *J. Geophys. Res. Solid Earth* **117**(B4), 1978–2012 (2012)
 - 3.15 F. Lyard, F. Lefèvre, T. Letellier, O. Francis: Modelling the global ocean tides: A modern insight from FES2004, *Ocean Dyn.* **56**, 394–415 (2006)
 - 3.16 R. Savcenko, W. Bosch: *EOT11a—Empirical Ocean Tide Model from Multi-Mission Satellite Altimetry*, Vol. 89 (Deutsches Geodätisches Forschungsinstitut, Munich 2010) p. 49
 - 3.17 M. Soffel: Report of the working group relativity for celestial mechanics and astrometry. In: *IAU Colloquium 180*, ed. by K. Johnston, D.D. McCarthy, B.J. Luzum, G.H. Kaplan (US Naval Observatory, Washington 2000) pp. 283–292
 - 3.18 A. Milani, A.M. Nobili, P. Farinella: *Non-Gravitational Perturbations and Satellite Geodesy* (Adam Hilger, Bristol 1987)
 - 3.19 G. Kopp, A. Fehlmann, W. Finsterle, D. Harber, K. Heuerman: Total solar irradiance data record accuracy and consistency improvements, *Metrologia* **49**, 29–33 (2012), S29–S33
 - 3.20 P.C. Knocke, J.C. Ries, B.D. Tapley: Earth radiation pressure effects on satellites, *Proc. AIAA/AAS Astrodyn. Conf. Minneapolis (AIAA, Reston 1988)* pp. 577–587
 - 3.21 C.J. Rodriguez-Solano, U. Hugentobler, P. Steigenberger, S. Lutz: Impact of earth radiation pressure on GPS position estimates, *J. Geod.* **86**(5), 309–317 (2012)
 - 3.22 D. Rubincam: LAGEOS orbit decay due to infrared radiation from earth, *J. Geophys. Res.* **92**, 1287–1294 (1987)
 - 3.23 R. Scharroo, K.F. Wakker, B.A.C. Ambrosius, R. Noomen: On the along-track acceleration of the LAGEOS satellite, *J. Geophys. Res.* **81**, 729–740 (1991)
 - 3.24 K.J. Sośnica: *Determination of Precise Satellite Orbits and Geodetic Parameters Using Satellite Laser Ranging*, Vol. 93 (Geodätisch-geophysikalische Arbeiten in der Schweiz, Schweizerische Geodätische Kommission, Zürich 2015)
 - 3.25 H.F. Fliegel, T.E. Gallini, E.R. Swift: Global positioning system radiation force model for geodetic applications, *J. Geophys. Res.* **97**(B1), 559–568 (1992)
 - 3.26 H.F. Fliegel, T.E. Gallini: Solar force modeling of block IIR global positioning system satellites, *J. Spacecr. Rockets* **33**(6), 863–866 (1996)
 - 3.27 W. Marquis, C. Krier: Examination of the GPS block IIR solar pressure model, *Proc. ION GPS 2000*, Salt Lake City, 2000 (Institute of Navigation, Virginia 2000) pp. 407–415
 - 3.28 M. Ziebart, P. Dare: Analytical solar radiation pressure modelling for GLONASS using a pixel array, *J. Geod.* **57**(11), 587–599 (2001)
 - 3.29 M. Ziebart: Generalized analytical solar radiation pressure modeling algorithm for spacecraft of complex shape, *J. Spacecr. Rockets* **41**(5), 840–848 (2004)
 - 3.30 M. Ziebart, S. Edwards, S. Adhya, A. Sibthorpe, P. Arrowsmith, P. Cross: High precision GPS IIR orbit prediction using analytical non-conservative force models, *Proc. ION GNSS 2004*, Long Beach, 2004 (ION, Virginia 2004) pp. 1764–1770
 - 3.31 M. Ziebart, S. Adhya, A. Sibthorpe, S. Edwards, P. Cross: Combined radiation pressure and thermal modelling of complex satellites: Algorithms and on-orbit tests, *Adv. Space Res.* **36**(3), 424–430 (2005)
 - 3.32 J.M. Dow, R.E. Neilan, C. Rizos: The international GNSS service in a changing landscape of global navigation satellite systems, *J. Geod.* **83**(3/4), 191–198 (2009)
 - 3.33 G. Beutler, E. Brockmann, W. Gurtner, U. Hugentobler, L. Mervart, M. Rothacher, A. Verdun: Extended orbit modeling techniques at the CODE processing center of the international GPS service for geodynamics (IGS): Theory and initial results, *Manuscr. Geod.* **19**, 367–386 (1994)
 - 3.34 A. Springer: *Modeling and Validating Orbits and Clocks Using the Global Positioning System*, Vol. 60 (Geodätisch-geophysikalische Arbeiten in der Schweiz, Schweizerische Geodätische Kommission, Zürich 1999)
 - 3.35 T. Springer, G. Beutler, M. Rothacher: A new solar radiation pressure model for GPS satellites, *GPS Solutions* **2**(3), 50–62 (1999)
 - 3.36 Y.E. Bar-Sever, K.M. Russ: *New and Improved Solar Radiation Models for GPS Satellites Based on Flight Data*, Final Report Task Plan 80–4193 (Jet Propulsion Laboratory, Pasadena 1997)
 - 3.37 Y.E. Bar-Sever, D. Kuang: *New Empirically Derived Solar Radiation Pressure Model for Global Positioning System Satellites*, IPN Progress Report 42–159 (Jet Propulsion Laboratory, Pasadena 2004)
 - 3.38 A. Sibthorpe, W. Bertiger, S.D. Desai, B. Haines, N. Harvey, J.P. Weiss: An evaluation of solar radiation pressure strategies for the GPS constellation, *J. Geod.* **85**(8), 505–517 (2011)
 - 3.39 C.J. Rodriguez-Solano, U. Hugentobler, P. Steigenberger: Adjustable box-wing model for solar radiation pressure impacting GPS satellites, *Adv. Space Res.* **49**(7), 1113–1128 (2012)
 - 3.40 C.J. Rodriguez-Solano, U. Hugentobler, P. Steigenberger, G. Allende-Alba: Improving the orbits of GPS block IIA satellites during eclipse seasons, *Adv. Space Res.* **52**(8), 1511–1529 (2013)
 - 3.41 C.J. Rodriguez-Solano, U. Hugentobler, P. Steigenberger, M. Bloßfeld, M. Fritsche: Reducing the draconitic errors in GNSS geodetic products, *J. Geod.* **88**, 559–574 (2014)
 - 3.42 M. Meindl, G. Beutler, D. Thaller, R. Dach, A. Jäggi: Geocenter coordinates estimated from GNSS data as viewed by perturbation theory, *Adv. Space Res.* **51**(7), 1047–1064 (2013)
 - 3.43 D. Arnold, M. Meindl, G. Beutler, R. Dach, S. Schaer, S. Lutz, L. Prange, K. Sosnica, L. Mervart, A. Jäggi: CODE's new solar radiation pressure model for GNSS orbit determination, *J. Geod.* **89**(8), 775–791 (2015)
 - 3.44 P. Steigenberger, U. Hugentobler, S. Loyer, F. Perosanz, L. Prange, R. Dach, M. Uhlemann, G. Gendt, O. Montenbruck: Galileo orbit and clock quality of the IGS multi-GNSS experiment, *Adv. Space Res.* **55**(1), 269–281 (2014)

- 3.45 O. Montenbruck, P. Steigenberger, U. Hugentobler: Enhanced solar radiation pressure modeling for galileo satellites, *J. Geod.* **89**(3), 283–297 (2015)
- 3.46 M. Ziebart, A. Sibthorpe, P. Cross, Y. Bar-Sever, B. Haines: Cracking the GPS-SLR orbit anomaly, *Proc. ION GNSS 2007*, Fort Worth (ION, Virginia 2007) pp. 2033–2038
- 3.47 Y. Vigue, B.E. Schutz, P.A.M. Abusali: Thermal force modeling for global positioning system satellites using the finite element method, *J. Spacecr. Rockets* **31**(5), 855–859 (1994)
- 3.48 S. Adhya, M. Ziebart, A. Sibthorpe, P. Arrowsmith, P. Cross: Thermal force modeling for precise prediction and determination of spacecraft orbits, *Navigation* **52**(3), 131–144 (2005)
- 3.49 J. Duha, G.B. Afonso, L.D. Damasceno Ferreira: Thermal re-emission effects on GPS satellites, *J. Geod.* **80**(12), 665–674 (2006)
- 3.50 U. Hugentobler: *Astrometry and Satellite Orbits: Theoretical Consideration and Typical Applications*, Vol. 57 (Geodätisch-geophysikalische Arbeiten in der Schweiz, Schweizerische Geodätische Kommission, Zürich 1998)
- 3.51 R.R. Allan, G.E. Cook: The long-period motion of the plane of a distant circular orbit, *Proc. R. Soc. Lond. A* **280**, 97–109 (1964)
- 3.52 A. Rossi: Resonant dynamics of medium earth orbits: Space debris issues, *Celest. Mech. Dyn. Astron.* **100**(4), 267–286 (2008)
- 3.53 B. Schutz, G. Giacaglia: Decade-scale gps orbit evolution and third body perturbations, *Proc. AIAA/AAS Astrodyn. Specialist Conf. Exhib. Honolulu* (AIAA, Reston 2008), AIAA 2008–7070
- 3.54 F. Deleflie, A. Rossi, C. Portmann, G. Métris, F. Barlier: Semi-analytical investigations of the long-term evolution of the eccentricity of Galileo and GPS-like orbits, *Adv. Space Res.* **47**(5), 811–821 (2011)
- 3.55 M. Fritsche, K. Sośnica, C.J. Rodríguez-Solano, P. Steigenberger, K. Wang, R. Dietrich, R. Dach, U. Hugentobler, M. Rothacher: Homogeneous re-processing of GPS, GLONASS and SLR observations, *J. Geod.* **88**(7), 625–642 (2014)
- 3.56 R. Jehn, A. Rossi, T. Flohrer, D. Navarro-Reyes: Re-orbiting of satellites in high altitudes, *Proc. 5th Eur. Conf. Space Debris*, Darmstadt (ESA, Noordwijk 2009)
- 3.57 C.C. Chao, R.A. Gick: Long-term evolution of navigation satellite orbits: GPS/GLONASS/GALILEO, *Adv. Space Res.* **34**(5), 1221–1226 (2004)
- 3.58 A.B. Jenkin, J.P. McVey: Constellation and graveyard collision risk for several MEO disposal strategies, *Proc. 5th Eur. Conf. Space Debris*, Darmstadt (ESA, Noordwijk 2009)
- 3.59 A. Rossi, I. Anselmo, C. Pardini, R. Jehn: Effectiveness of the de-orbiting practices in the meo region, *Proc. 5th Eur. Conf. Space Debris*, Darmstadt (ESA, Noordwijk 2009)
- 3.60 J. Radtke, R. Domínguez-González, S. Flegel, N. Sánchez-Ortiz, K. Merz: Impact of eccentricity build-up and graveyard disposal strategies on MEO navigation constellations, *Adv. Space Res.* **56**(11), 2626–2644 (2015)
- 3.61 J. Griffiths, J. Ray: On the precision and accuracy of IGS orbits, *J. Geod.* **83**, 277–287 (2009)
- 3.62 M. Pearlman, J. Degnan, J. Bosworth: The international laser ranging service, *Adv. Space Res.* **30**(2), 125–143 (2002)
- 3.63 J.J. Miller, J. LaBrecque, A.J. Oria: Laser reflectors to ride on board GPS III, *GPS World* **24**(9), 12–17 (2013)
- 3.64 J. Griffiths, J.R. Ray: Sub-daily alias and draconitic errors in the IGS orbits, *GPS Solutions* **17**, 413–422 (2013)
- 3.65 C. Urschl, G. Beutler, W. Gurtner, U. Hugentobler, S. Schaer: Contribution of SLR tracking data to GNSS orbit determination, *Adv. Space Res.* **39**(10), 1515–1523 (2007)
- 3.66 K. Sośnica, D. Thaller, R. Dach, P. Steigenberger, G. Beutler, D. Arnold: Satellite laser ranging to GPS and GLONASS, *J. Geod.* **89**(7), 725–743 (2015)
- 3.67 P. Steigenberger, U. Hugentobler, A. Hauschild, O. Montenbruck: Orbit and clock analysis of compass GEO and IGS0 satellites, *J. Geod.* **87**(6), 515–526 (2013)
- 3.68 K. Chen, T. Xu, G. Chen, J. Li, S. Yu: The orbit and clock combination of iGMAS analysis centers and the analysis of their precision, *Proc. China Satell. Navig. Conf. (CSNC)*, Xi'an, Vol. 2, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin, Heidelberg 2015) pp. 421–438
- 3.69 Navstar GPS Space Segment/Navigation User Segment Interfaces, Interface Specification, IS-GPS-200H (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo 2013)
- 3.70 Global Navigation Satellite System GLONASS – Interface Control Document, v5.1 (Russian Institute of Space Device Engineering, Moscow 2008)
- 3.71 European GNSS (Galileo) Open Service Signal In Space Interface Control Document, OS SIS ICD, Iss. 1.2, Nov. 2015 (European Union 2015)
- 3.72 BeiDou Navigation Satellite System Signal in Space Interface Control Document – Open Service Signal (China Satellite Navigation Office, Beijing 2013)
- 3.73 Quasi-Zenith Satellite System Navigation Service Interface Specification for QZSS, IS-QZSS, v1.6, 28 Nov. 2014 (JAXA, Chofu 2014)
- 3.74 O. Montenbruck, P. Steigenberger, A. Hauschild: Broadcast versus precise ephemerides: A multi-GNSS perspective, *GPS Solutions* **19**(2), 321–333 (2015)
- 3.75 U. Rossbach: Positioning and Navigation Using the Russian Satellite System GLONASS, Ph.D. Thesis (Univ. d. Bundeswehr München, Neubiberg 2001)
- 3.76 H. Yin, Y. Morton, M. Carroll, E. Vinande: Performance analysis of L2 and L5 CNAV broadcast ephemeris for orbit calculation, *Proc. ION ITM 2014*, San Diego (ION, Virginia 2014) pp. 761–768
- 3.77 B.W. Remondi: Computing satellite velocity using the broadcast ephemeris, *GPS Solutions* **8**(3), 181–183 (2004)
- 3.78 J. Zhang, K. Zhang, R. Grenfell, R. Deakin: GPS satellite velocity and acceleration determination using the broadcast ephemeris, *J. Navig.* **59**(2), 293–305 (2006)
- 3.79 R. Marson, S. Lagrasta, F. Malvolti, T.S.V. Tiburtina: Fast generation of precision orbit ephemeris, *Proc. ION ITM 2011*, San Diego (ION, Virginia 2011) pp. 565–576

- 3.80 O. Montenbruck, P. Steigenberger: The BeiDou Navigation Message, *J. Glob. Position. Syst.* **12**(1), 1–12 (2013)
- 3.81 M. Stewart, M. Tsakiri: GLONASS broadcast orbit computation, *GPS Solutions* **2**(2), 16–27 (1998)
- 3.82 Minimum Operational Performance Standards for GPS/WAAS Airborne Equipment, RTCA DO-229D (RTCA, Washington DC 2006)
- 3.83 T. Reid, T. Walter, P. Enge: L1/L5 SBAS MOPS ephemeris message to support multiple orbit classes, *Proc. ION ITM*, San Diego (ION, Virginia 2013) pp. 78–92
- 3.84 A.J. Dorsey, W.A. Marquis, P.M. Fyfe, E.D. Kaplan, L.F. Wiederholt: GPS system segments. In: *Understanding GPS: Principles and Applications*, ed. by E.D. Kaplan, C.J. Hegarty (Artech House, Norwood 2006) pp. 67–112
- 3.85 R. DiEposti, J. DiLellio, C. Kelley, A. Dorsey, H. Fliegel, J. Berg, C. Edgar, T. McKendree, P. Shome: The proposed state vector representation of broadcast navigation message for user equipment implementation of GPS satellite ephemeris propagation, *Proc. ION NTM*, San Diego (ION, Virginia 2004) pp. 294–312
- 3.86 D.L. Warren, J.F. Raquet: Broadcast versus precise GPS ephemerides: A historical perspective, *GPS Solutions* **7**(3), 151–156 (2003)
- 3.87 L. Heng, G.X. Gao, T. Walter, P. Enge: Statistical characterization of GPS signal-in-space errors, *Proc. ION ITM*, San Diego (ION, Virginia 2011) pp. 312–319
- 3.88 L. Heng, G.X. Gao, T. Walter, P. Enge: Statistical characterization of GLONASS broadcast ephemeris errors, *Proc. ION GNSS 2011*, Portland (ION, Virginia 2011) pp. 3109–3117
- 3.89 O. Montenbruck, R. Schmid, F. Mercier, P. Steigenberger, C. Noll, R. Fatkulin, S. Kogure, S. Ganeshan: GNSS Satellite Geometry and Attitude Models, *Adv. Space Res.* **56**(6), 1015–1029 (2015)
- 3.90 J. Kouba: A simplified yaw-attitude model for eclipsing GPS satellites, *GPS Solutions* **13**(1), 1–12 (2009)
- 3.91 F. Dilssner: GPS IIF-1 satellite, antenna phase center and attitude modeling, *Inside GNSS* **5**(6), 59–64 (2010)
- 3.92 F. Dilssner, T. Springer, W. Enderle: GPS IIF yaw attitude control during eclipse season, *Proc. Am. Geophys. Union Fall Meet.*, San Francisco (AGU, Washington 2011)
- 3.93 F. Dilssner, T. Springer, G. Gienger, J. Dow: The GLONASS-M satellite yaw-attitude model, *Adv. Space Res.* **47**(1), 160–171 (2011)
- 3.94 Y. Ishijima, N. Inaba, A. Matsumoto, K. Terada, H. Yonechi, H. Ebisutani, S. Ukava, T. Okamoto: Design and development of the first quasi-zenith satellite attitude and orbit control system, *Proc. IEEE Aerosp. Conf.* (2009) pp. 1–8
- 3.95 W. Wang, G. Chen, S. Guo, X. Song, Q. Zhao: A study on the Beidou IGS0/MEO satellite orbit determination and prediction of the different yaw control mode, *Proc. China Satell. Navig. Conf. (CSNC) 2013*, Wuhan, Vol. III, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin Heidelberg 2013) pp. 31–40
- 3.96 J. Guo, Q. Zhao, T. Geng, X. Su, J. Liu: Precise orbit determination for COMPASS IGS0 satellites during Yaw maneuvers, *Proc. CSNC*, Wuhan, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin, Heidelberg 2013) pp. 41–53, Vol. III
- 3.97 A. Hauschild, P. Steigenberger, C. Rodriguez-Solano: QZS-1 yaw attitude estimation based on measurements from the CONGO network, *Navigation* **59**(3), 237–248 (2012)
- 3.98 J. Guo, Q. Zhao: Analysis of precise orbit determination for BeiDou satellites during yaw maneuvers, *Proc. CSNC*, Wuhan (2014)
- 3.99 X. Dai, M. Ge, Y. Lou, C. Shi, J. Wickert, H. Schuh: Estimating the yaw-attitude of BDS IGS0 and MEO satellites, *J. Geod.* **89**(10), 1005–1018 (2015)

Signals and Modulation

Michael Meurer, Felix Antreich

Satellite navigation relies on signals radiated by orbiting satellites and received by mobile satellite navigation receivers. This chapter addresses the fundamentals of such navigation signals and introduces the most important underlying concepts. It provides an introduction to radio frequency signals including the basics of electromagnetic waves, their carrier frequency, polarization, as well as group and phase velocity. The application of waves for carrying signals, their power and spectrum are addressed. It is shown how information-carrying signals can be modulated onto the wave using various modulation schemes such as binary phase shift keying, binary offset carrier, and alternating binary offset carrier. Setting out from international agreed allocations, the frequency bands used in GNSS are described. The concept of pseudo-random codes which is typically used for GNSS signals is introduced as well as their receiver side processing following a correlation principle.

4.1	Radiofrequency Signals	91
4.1.1	Maxwell's Theory of Electromagnetic Waves and Electromagnetic Foundation	91
4.1.2	Modulation and Complex Baseband Representation of Signals	93
4.1.3	Frequency Bands and Polarization	96
4.2	Spread Spectrum Technique and Pseudo Random Codes	97
4.2.1	Spread Spectrum Signals for Ranging	97
4.2.2	Pseudo-Random Binary Sequences	99
4.2.3	Correlation and Time-Delay Estimation	102
4.3	Modulation Schemes	107
4.3.1	Binary Phase Shift Keying	107
4.3.2	Binary Offset Carrier Modulation and Derivatives	109
4.4	Signal Multiplexing	113
4.4.1	Interplex	114
4.4.2	AltBOC	116
4.5	Navigation Data and Data-Free Channels	117
	References	118

4.1 Radiofrequency Signals

4.1.1 Maxwell's Theory of Electromagnetic Waves and Electromagnetic Foundation

The propagation of navigation signals through space relies on electromagnetic waves. In this section, a brief introduction to wave propagation is given as needed in the scope of this book. For a further detailed introduction to electromagnetic theory, the reader is referred to [4.1].

The physical foundation of all electromagnetic phenomena is Maxwell's theory and its corresponding equations [4.1–6]. Using \mathbf{E} and \mathbf{H} as three-dimensional vectors describing the electric and the magnetic field strength, respectively, Maxwell's equations take the fol-

lowing coordinate independent form

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}, \quad (4.1)$$

$$\nabla \times \mathbf{H} = \varepsilon \frac{\partial \mathbf{E}}{\partial t}, \quad (4.2)$$

$$\nabla \cdot \mathbf{E} = 0, \quad (4.3)$$

$$\nabla \cdot \mathbf{H} = 0. \quad (4.4)$$

The variable t denotes time, whereas the constants ε and μ are the electrical permittivity and the magnetic permeability, respectively. The notations $\nabla \times$ and $\nabla \cdot$ denote the curl and divergence operators for vectorial fields [4.7]. By further applying the curl operation

on (4.1), we obtain

$$\begin{aligned}\nabla \times (\nabla \times \mathbf{E}) &= \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}, \\ &= -\mu \frac{\partial}{\partial t} \nabla \times \mathbf{H},\end{aligned}\quad (4.5)$$

where ∇ and ∇^2 denote the gradient and vectorial Laplacean operators. Combining (4.2) with (4.3) and (4.5) yields the fundamental wave equations

$$\nabla^2 \mathbf{E} = \varepsilon \mu \frac{\partial^2}{\partial t^2} \mathbf{E}, \quad (4.6)$$

$$\nabla^2 \mathbf{H} = \varepsilon \mu \frac{\partial^2}{\partial t^2} \mathbf{H}. \quad (4.7)$$

The solutions of (4.6) and (4.7) describe electromagnetic waves which propagates through space.

For further discussion, let us consider a Cartesian coordinate system with coordinates x, y, z , the position vector $\mathbf{r} = (x, y, z)^\top$ and the corresponding components of the electric field vector $\mathbf{E} = (E_x, E_y, E_z)^\top$. Moreover, for a moment let us assume that the electrical field vector \mathbf{E} is not a function of x and y . In this case, all partial derivatives in (4.6) with respect to x and y vanish, that is,

$$\frac{\partial \mathbf{E}}{\partial x} = \frac{\partial \mathbf{E}}{\partial y} = 0. \quad (4.8)$$

Solving (4.6) and using (4.8) it can be shown that

$$\mathbf{E}(z, t) = \mathbf{E}_0 \cos \left[2\pi f_c \left(\frac{z}{v_p} - t \right) \right] \quad (4.9)$$

is a solution of the fundamental wave equations of (4.6), where

$$v_p = \frac{1}{\sqrt{\varepsilon \mu}}. \quad (4.10)$$

is the propagation speed or phase velocity of the wave and f_c its carrier frequency. In free space, v_p is identical to the speed of light c . Equation (4.9) describes a sinusoidal electromagnetic wave propagating in the direction of the positive z -axis. $\mathbf{E}(z, t)$ is a periodic function in t and z , where the spatial period is quantified by the wavelength

$$\lambda = \frac{v_p}{f_c}. \quad (4.11)$$

The wavelength λ is the distance traveled by the wave within one period $1/f_c$ of the wave. Substituting (4.11) into (4.9) yields

$$\mathbf{E}(z, t) = \mathbf{E}_0 \cos \left[\underbrace{\frac{2\pi}{\lambda}}_{=k} (z - ct) \right], \quad (4.12)$$

where k is termed wave number. The result of (4.12) can easily be generalized to waves propagating in any direction characterized by the wave vector

$$\mathbf{k} = \frac{2\pi}{\lambda} \mathbf{n}_0, \quad (4.13)$$

where \mathbf{n}_0 is a unit vector pointing into the direction of propagation. Using (4.13) one obtains

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 \cos \left(\mathbf{k}^\top \mathbf{r} - \frac{2\pi}{\lambda} c t + \varphi_0 \right), \quad (4.14)$$

where φ_0 is an additional degree of freedom termed zero-phase offset. At each location within a plane which is perpendicular to \mathbf{k} , that is, in all locations in which $\mathbf{k}^\top \mathbf{r}$ is constant, the same electrical field strength can be observed. Therefore, the wave characterized by (4.14) is a planar wave. If further \mathbf{E}_0 is orthogonal to \mathbf{k} , the wave is a transversal planar wave. Usually electromagnetic waves which are relevant in satellite navigation can be considered to be of that kind.

So far, we only discussed the electrical field strength \mathbf{E} . However, similar considerations can be made for the magnetic field strength \mathbf{H} . Moreover, it can be shown that in a homogenous, isotropic, and stationary medium, the electrical field strength \mathbf{E} and the magnetic field strength \mathbf{H} are always orthogonal to each other, while both of them are orthogonal to the propagation direction described by the wave vector \mathbf{k} (Fig. 4.1).

For the electromagnetic wave of (4.14) the ratio between the components of the electrical field strength \mathbf{E} is constant and defined by \mathbf{E}_0 . Such a wave is termed linearly polarized, see Fig. 4.1. If two linearly polarized waves with identical wave vector \mathbf{k} but orthogonal $\mathbf{E}_{0,1}$ and $\mathbf{E}_{0,2}$ are superposed, the electrical field strength $\mathbf{E}(\mathbf{r}, t)$ at a certain location \mathbf{r} experiences in general over time an elliptical variation. If further the absolute magnitudes of $\mathbf{E}_{0,1}$ and $\mathbf{E}_{0,2}$ are identical and the zero-phase offsets $\varphi_{0,1}$ and $\varphi_{0,2}$ differ by $\pi/2$, the elliptical variation gets a circular one. Waves of that type are termed circularly polarized (see Fig. 4.1, right). If the electric field vector rotates clockwise, when looking into the direction of propagation, the electromagnetic wave is right-handed polarized, if not it is left handed.

Electromagnetic waves propagating through ionized gases or through the Earth magnetic field undergo changes in their polarization causing linearly polarized waves to become elliptically or circularly polarized [4.8]. All current satellite navigation systems circumvent this effect by using right-handed circularly polarized (RHCP) signals.

The Austrian physicist Christian Doppler postulated in 1842 that a relative motion between transmitter and

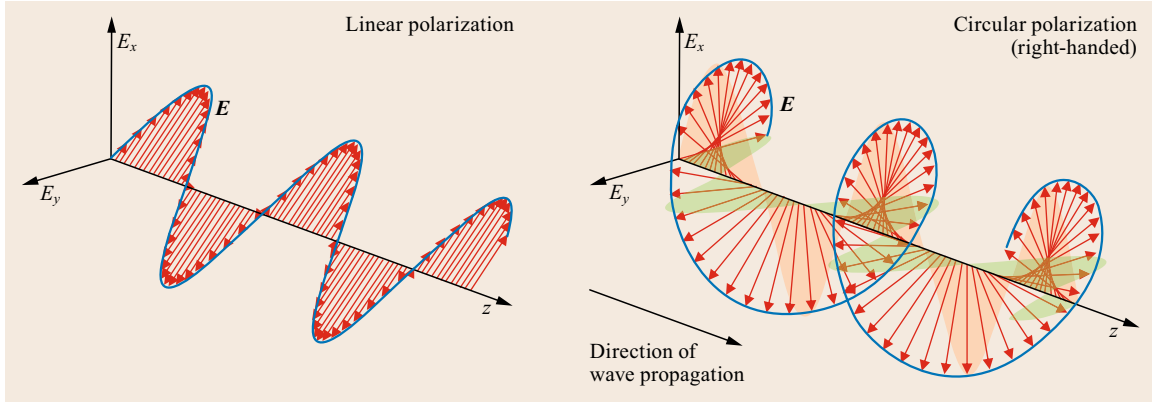


Fig. 4.1 Linear and circular polarization of radio waves

receiver of a wave will cause a frequency shift, which is today known as Doppler shift or Doppler effect [4.8]. Let us shortly revisit our previous simplified example of (4.9) and assume that the transmitter of the radio wave is moving into the direction of the positive z -axis with velocity v . Moreover, the origin of the coordinate system is assumed to be at the (moving) location of the transmitter (Fig. 4.2). In this case, the (time dependent) position vector $\mathbf{r}_{\text{rx}}(t)$ of a stationary receiver located on the z -axis gets

$$\mathbf{r}_{\text{rx}}(t) = \begin{pmatrix} 0 \\ 0 \\ r_{\text{rx},0} \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ v \end{pmatrix} t. \quad (4.15)$$

Applying the position vector of (4.15) onto the wave equation (4.9) yields the oscillation of the electric field strength observed by the receiver

$$\begin{aligned} E(\underbrace{r_{\text{rx},0} - vt}_{=z}, t) \\ &= E_0 \cos \left[2\pi f_c \left(\frac{r_{\text{rx},0} - vt}{c} - t \right) \right] \\ &= E_0 \cos \left(2\pi \left[f_c \frac{r_{\text{rx},0}}{c} - \underbrace{\left(1 + \frac{v}{c} \right) f_c t}_{=f_c + f_D} \right] \right). \end{aligned} \quad (4.16)$$

The quantity

$$f_D = \frac{v}{c} f_c \quad (4.17)$$

is termed Doppler shift or Doppler frequency and describes the frequency shift experienced by the receiver. Introducing the unit vector \mathbf{n} pointing from the transmitter to the receiver and the velocity vector \mathbf{v} of the transmitter relative to the receiver, the result of (4.17) can be generalized to

$$f_D = \frac{\mathbf{v}^\top \mathbf{n}}{c} f_c. \quad (4.18)$$

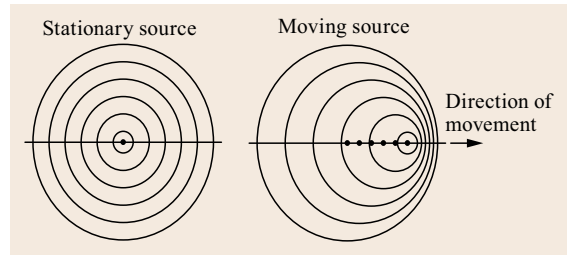


Fig. 4.2 Doppler effect for a moving source of a radio wave

4.1.2 Modulation and Complex Baseband Representation of Signals

In the following, signals are described in time domain. Doing so, signals are denoted by lower case italic letters and are written as functions of the time t , for example, $x(t)$. Figure 4.3a depicts a signal example. Alternatively, signals can be described in frequency domain as functions of the frequency f . Signals in frequency domain are denoted by upper case italic letters, for example, $X(f)$. $X(f)$ is termed spectrum of $x(t)$ and is achieved by Fourier transformation

$$X(f) = \mathcal{F}\{x(t)\} = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt. \quad (4.19)$$

of the time domain signal $x(t)$. Figure 4.3b shows a corresponding example of a spectrum $X(f)$.

Let us consider an incoming electromagnetic wave which leads at the receiver to the received radio frequency (RF) signal of the form

$$x_{\text{RF}}(t) = d_c(t) \cos[2\pi f_c(t) t + \varphi_c(t)], \quad (4.20)$$

where $d_c(t)$, $f_c(t)$, and $\varphi_c(t)$ define the amplitude, frequency, and phase of the harmonic signal, respectively.

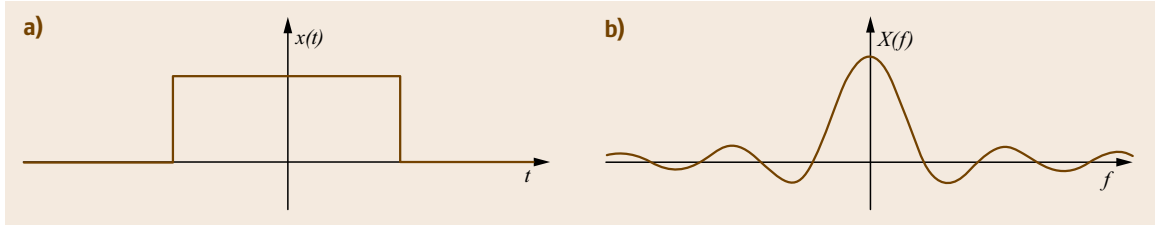


Fig. 4.3a,b Example of a signal in (a) time domain and (b) frequency domain

In case of constant parameters $d_c(t)$, $f_c(t)$, and $\varphi_c(t)$, the signal $x_{\text{RF}}(t)$ is a pure sinusoidal carrier which does not carry any information. However, it is desired in satellite navigation as well as communications to put additional information onto the signal, for example, in order to transfer navigation data from the transmitter to the receiver or simply to distinguish one signal from the other. By varying one or several of the aforementioned three parameters over time, additional information can be put onto the signal. This process is termed modulation. Depending on whether $d_c(t)$, $f_c(t)$, or $\varphi_c(t)$ are varied over time, we distinguish between amplitude modulation, frequency modulation, and phase modulation. Let us now consider the case where only $d_c(t)$ and $\varphi_c(t)$ vary over time, that is,

$$x_{\text{RF}}(t) = d_c(t) \cos(2\pi f_c t + \varphi_c(t)) \quad (4.21)$$

holds for the received signal with constant f_c . $d_c(t)$ is termed envelope or baseband signal. In Sect. 4.2 it will be shown how the baseband signal $d_c(t)$ has to be chosen in order to carry data as well as further structural information (termed spreading code for uniquely identifying signals in satellite navigation). Using the trigonometric identity

$$\cos(\alpha + \beta) = \cos(\alpha) \cos(\beta) - \sin(\alpha) \sin(\beta), \quad (4.22)$$

the RF signal (4.21) can be reformulated as

$$\begin{aligned} x_{\text{RF}}(t) &= \underbrace{d_c(t) \cos(\varphi_c(t))}_{=\sqrt{2}x_I(t)} \cos(2\pi f_c t) \\ &\quad - \underbrace{d_c(t) \sin(\varphi_c(t))}_{=\sqrt{2}x_Q(t)} \sin(2\pi f_c t) \\ &= +\sqrt{2}x_I(t) \cos(2\pi f_c t) - \sqrt{2}x_Q(t) \sin(2\pi f_c t). \end{aligned} \quad (4.23)$$

The baseband signals $x_I(t)$ and $x_Q(t)$ are termed inphase and quadrature component of the RF signal $x_{\text{RF}}(t)$. The factors $\sqrt{2}$ are introduced for power normalization as shown in [4.9]. The process of modulating an in-phase and quadrature component onto the carrier signal is termed quadrature amplitude modulation (QAM) and is visualized in Fig. 4.4.

It is compact and convenient to write QAM radio signals as

$$x_{\text{RF}}(t) = \sqrt{2} \text{Re}\{\tilde{x}(t) e^{j2\pi f_c t}\}. \quad (4.24)$$

The complex-valued signal

$$\tilde{x}(t) = x_I(t) + jx_Q(t) \quad (4.25)$$

is termed complex baseband signal or complex envelope. The notation of complex baseband signals provides the basis for the following sections.

Let us shortly revisit the monochromatic sinusoidal wave (4.12) moving in the positive z -direction, which can be rewritten in a slightly modified form as

$$E(z, t) = E_0 \cos(kz - 2\pi f_c t). \quad (4.26)$$

As z or t change, so does the phase $kz - 2\pi f_c t$ of the wave. The phase is constant if $kz - 2\pi f_c t$ is constant, that is, if

$$z = \frac{2\pi f_c}{k} t = v_p t, \quad (4.27)$$

where v_p is the phase velocity introduced in (4.10). As the wave propagates, the whole sinusoidal pattern moves with velocity v_p toward the positive z -axis [4.10].

Now let us consider a simple amplitude-modulated signal

$$x_{\text{RF}}(t) = \cos(2\pi f_d t) \cos(2\pi f_c t), \quad (4.28)$$

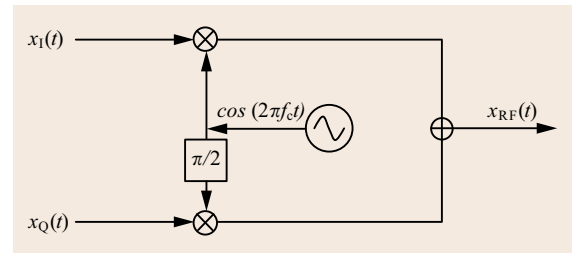


Fig. 4.4 Quadrature modulation with in-phase and quadrature components $x_I(t)$ and $x_Q(t)$

where $\cos(2\pi f_d t)$ describes the baseband message signal with modulation frequency $f_d \ll f_c$, modulated onto the carrier signal. The signal is visualized in Fig. 4.5a. Applying trigonometric identities, (4.28) may also be written as

$$x_{\text{RF}}(t) = \frac{1}{2} (\cos[2\pi(f_c + f_d)t] + \cos[2\pi(f_c - f_d)t]) . \quad (4.29)$$

Obviously, the modulated signal $x_{\text{RF}}(t)$ is the sum of two sinusoidal components with slightly different frequencies $f_c + f_d$ and $f_c - f_d$.

Now let us further assume that the signal described by (4.29) is the source of an electromagnetic wave (4.26), which propagates through a dispersive medium. A dispersive medium is characterized by the fact that dielectric permittivity ϵ and/or the magnetic permeability μ , and therefore the phase velocity v_p of (4.10), are frequency dependent. Consequently, both sinusoidal components of (4.29) propagate with slightly different phase velocities and have slightly different wavelengths λ and wave numbers k .

Let the wave number for a carrier of frequency f_c be k and the wave numbers corresponding to the slightly

deviating frequencies $f_c + f_d$ and $f_c - f_d$ be $k + \Delta k$ and $k - \Delta k$, respectively. Then we obtain for the electromagnetic wave after traveling a distance of z from the origin along the z -axis

$$E(z, t) = \frac{1}{2} E_0 (\cos[(k + \Delta k)z - 2\pi(f_c + f_d)t] + \cos[(k - \Delta k)z - 2\pi(f_c - f_d)t]) . \quad (4.30)$$

Applying the same trigonometric identity used earlier, (4.30) can be rewritten as

$$\begin{aligned} E(z, t) &= E_0 \cos(\Delta k z - 2\pi f_d t) \cos(kz - 2\pi f_c t) \\ &= E_0 \cos \left[2\pi f_d \left(t - \frac{1}{2\pi} \frac{\Delta k}{f_d} z \right) \right] \\ &\quad \times \cos \left[2\pi f_c \left(t - \frac{1}{2\pi} \frac{k}{f_c} z \right) \right] . \end{aligned} \quad (4.31)$$

As f_d gets smaller, $\Delta k/f_d$ approaches dk/df [4.10] and the electromagnetic wave can be expressed as

$$\begin{aligned} E(z, t) &= E_0 \cos \left[2\pi f_d \left(t - \frac{1}{2\pi} \frac{dk}{df} z \right) \right] \\ &\quad \times \cos \left[2\pi f_c \left(t - \frac{1}{2\pi} \frac{k}{f_c} z \right) \right] . \end{aligned} \quad (4.32)$$

This result implies that the carrier and the modulation of a electromagnetic wave in a dispersive medium propagate at different speeds. The carrier travels with the phase velocity v_p of (4.10) and the modulation, that is, the baseband message signal $\cos(2\pi f_d t)$, propagates with the group velocity

$$v_g = 2\pi \left. \frac{df}{dk} \right|_{f=f_c} . \quad (4.33)$$

Using (4.10) and (4.33), we can reformulate (4.32) to obtain the alternative expression

$$\begin{aligned} E(z, t) &= E_0 \cos \left[2\pi f_d \left(t - \frac{z}{v_g} \right) \right] \\ &\quad \times \cos \left[2\pi f_c \left(t - \frac{z}{v_p} \right) \right] . \end{aligned} \quad (4.34)$$

The consequence of the difference between phase velocity and group velocity is a divergence in propagation between carrier and modulation of an electromagnetic wave. In Fig. 4.5b, this effect is well observable. This phenomenon is referred to as code-carrier divergence in satellite navigation [4.10].

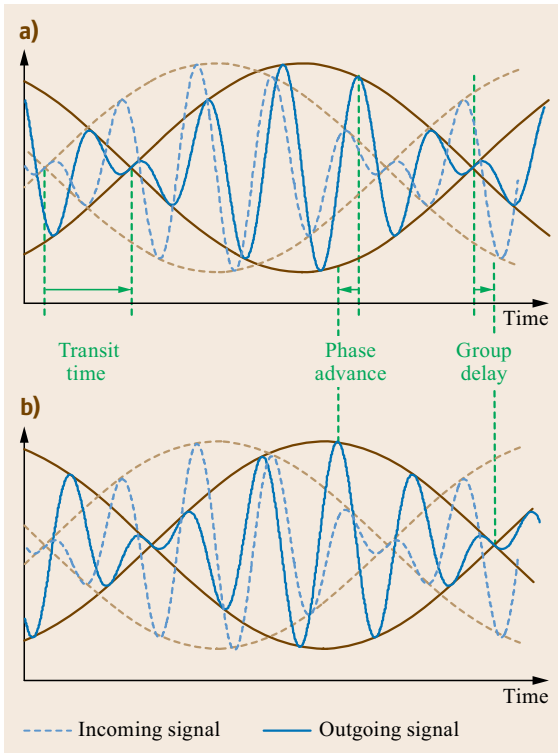


Fig. 4.5a,b Propagation of a modulated signal in (a) non-dispersive and (b) dispersive medium

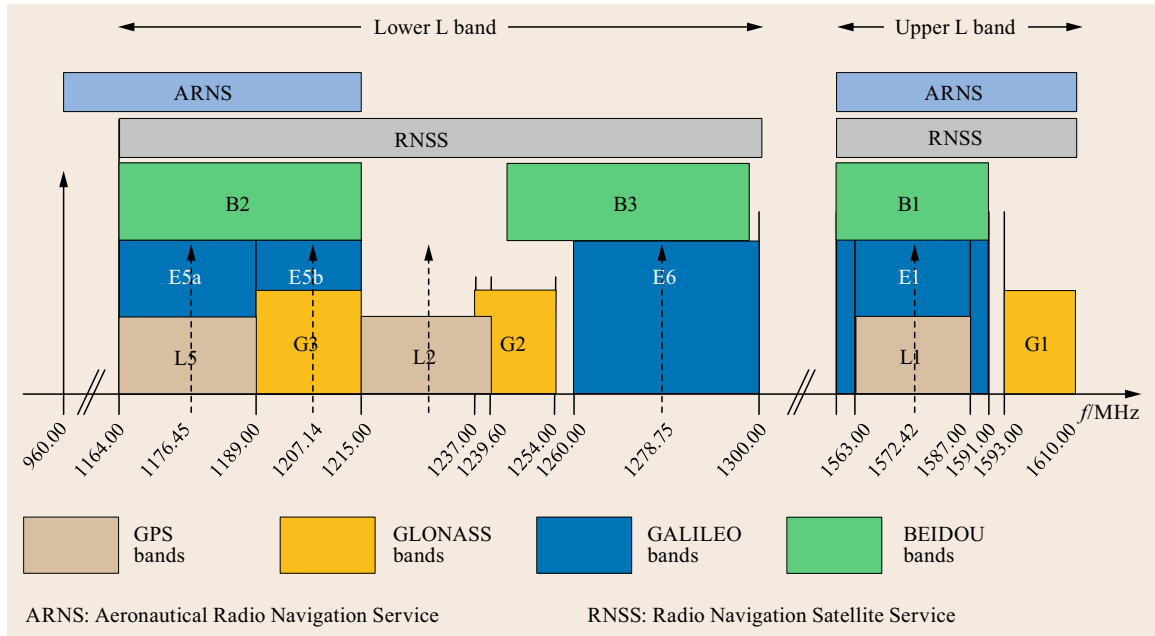


Fig. 4.6 Frequency bands used by global satellite navigation systems – ITU frequency allocations

4.1.3 Frequency Bands and Polarization

The carrier frequency influences many aspects of satellite navigation signals. This reaches from the propagation behaviors including Doppler effect over interference environment to aspects of necessary hardware components and related imperfections. Today's global satellite navigation systems all make use of the frequency band between 1 and 2 GHz which is termed L-band. In the L-band several subfrequency bands have been allocated and made available by the International Telecommunications Union (ITU). The L-band offers several advantages for the usage in satellite navigation: the propagation conditions are quite good including especially moderate attenuation and impact of atmospheric effects, rain etc. The antenna size for signals in L-band is quite limited which is the basis for miniaturized satellite navigation receivers and mobile applications. Moreover, a large variety of mature hardware components is available at low costs.

Two different swaths of spectrum, the upper L-band (1559–1610 MHz) and the lower L-band (1164–1300 MHz) have been established for worldwide satellite navigation use. This portion of spectrum is termed RNSS (radio navigation satellite service) band. For safety-critical applications such as aeronautical applications a certain protection of the navigation signals

against interferences is necessary. For such services, the ARNS (aeronautical radio navigation service) band is allocated which offers specific protection against such threats (see also Chap. 16). Figure 4.6 gives an overview of the available frequency bands as filed at ITU also showing the RNSS and ARNS band allocations. Details on the individual usage of the subfrequency bands by different global systems can be found in Annex B. The use of two or more distinct carrier frequencies within one satellite navigation system service offers several advantages and improved receiver performance. This includes robustness aspects as well as the possibility to mitigate ionospheric propagation errors.

Global satellite navigation systems consistently use right-handed circularly polarized waves (Sect. 4.1.1) for the transmitted signals. The circular polarization of GNSS signals is preferred over linear polarization to limit possible losses caused by orientation mismatch between the incident electromagnetic field and the receiving antenna. The exclusive choice of right-handed polarization (which was first adopted by GPS) in all other GNSSs is motivated by the intention to improve interoperability among different systems. It ensures that a user antenna capable of receiving right-handed polarized signals can be jointly used with a multitude of different satellite navigation systems.

4.2 Spread Spectrum Technique and Pseudo Random Codes

In this section, the main properties of spread spectrum signals are presented. We explore the motivation for GNSS to use spread spectrum signals and we discuss the use of pseudo-random (PR) binary sequences which are used as spreading codes to spread GNSS signals.

It is shown that design of GNSS signals can be separated into two problems, that is, the design of PR binary sequences and the design of chip pulse shapes which are convolved with these sequences. Both, the PR binary sequences and the respective chip pulse shapes define the performance of synchronization parameter estimation, that is, time-delay, carrier phase, and Doppler frequency, and consequently the performance of pseudo range as well as position estimation.

Basic principles of correlation and time-delay estimation are presented. It is shown how a standard GNSS receiver is performing time-delay estimation and based on this, important properties of PR binary sequences and chip pulse shapes are discussed and their influence on synchronization parameter estimation performance is demonstrated.

4.2.1 Spread Spectrum Signals for Ranging

Several satellites need to share the same transmission medium and broadcast to the GNSS users in order to enable positioning. The satellites need to share the transmission medium such that the GNSS users can separate different satellites, perform ranging, and receive the navigation data. The satellites need to share the available bandwidth by using propagation channel access or multiple access (MA) techniques. In general there are three basic MA techniques:

- Time division multiple access (TDMA)
- Frequency division multiple access (FDMA)
- Code division multiple access (CDMA).

These three MA techniques separate signals in time, in frequency or by codes with the aim to ensure that

the signals will be separated or even orthogonal. Each of these techniques can achieve, in principle, the same aggregate spectral efficiency, as they achieve the same symbol rate per channel user (satellite), the same number of channel users (satellites), and the same total bandwidth. However, each scheme has advantages and disadvantages. These basic MA techniques can also be combined to form hybrid combinations, for example, frequency division and time division (FD/TDMA) or combined frequency division and code division (FD/CDMA), and others [4.11].

The principle of TDMA, FDMA, and CDMA is illustrated in Fig. 4.7, where different shaded blocks resemble different channel users (satellites). In the case of TDMA (Fig. 4.7a), channel users occupy the complete available bandwidth but at different time slots; thus, they transmit in turns in assigned time slots. In the case of FDMA (Fig. 4.7b), channel users all transmit at the same time, but they occupy different sub-bands of the total available bandwidth. In the case of CDMA (Fig. 4.7c), all channel users transmit at the same time and use the complete available bandwidth, but the channel users are separated by codes.

Since for GNSS, the number of in-view satellites (channel users) is quite low (around maximum 12 per system) and data transmission demands in general do not play a prominent role, mainly other performance measures have to be considered when choosing the appropriate MA technique, this includes, for example:

- Synchronization accuracy
- Intersystem multiple access interference (MAI-R) or spectral separation
- Intrasystem multiple access interference (MAI-A)
- Interference robustness
- Multipath performance
- Signal flexibility
- Bandwidth efficiency
- Implementation issues etc.

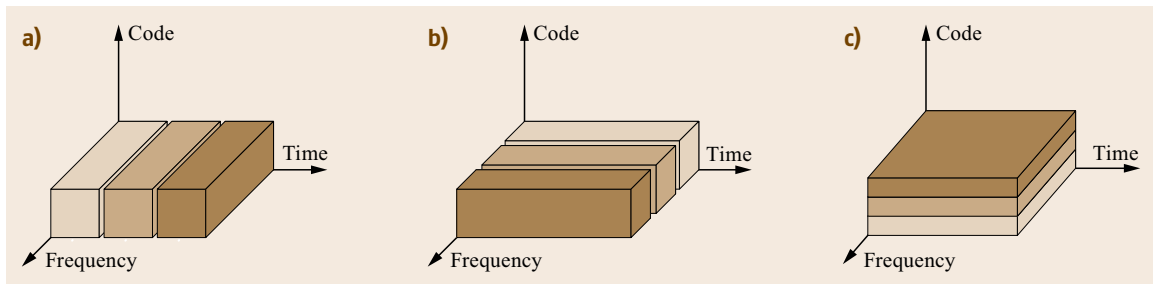


Fig. 4.7 (a) Time division multiple access (TDMA), (b) frequency division multiple access (FDMA), and (c) code division multiple access (CDMA)

For GNSSs, either CDMA or FDMA and when considering spectral separation between GNSSs, either FD/CDMA or CD/FDMA concepts are applied in order to establish MA for different satellites of different GNSSs with respect to the allocated time–frequency resources. Most GNSSs basically use direct sequence CDMA (DS-CDMA) [4.11] for which the signal energy of each satellite is continuously distributed throughout the entire time–frequency plane. The different satellites transmit their signals in the same frequency band at the same time. Each satellite uses a different code for transmitting its signal. Spectral separation between different GNSSs in the same frequency band can be achieved by usage of different modulation schemes. The GLONASS system uses FDMA where each visible satellite uses a different frequency slot and also uses a code sequence for synchronization and channel parameter estimation purposes.

We assume phase coherent frequency down-conversion of the radio frequency signal to baseband of either considering the carrier frequency with respect to a DS-CDMA system (e.g., GPS, Galileo, and Beidou; see Chapters 7, 9, 10) or each carrier of the respective satellite of a CD/FDMA system (e.g., GLONASS; Chap. 8). This implies that the carrier phase was estimated without error and the Doppler effect could be compensated perfectly. Baseband means that the frequency support of the signal is centered around the origin at 0 Hz. Thus, the received DS-CDMA or CD/FDMA baseband signal of one satellite can be given as

$$y(t) = \sqrt{P} m(t) c(t - \tau) + n(t), \quad (4.35)$$

where P denotes the signal power, $c(t)$ is the PR spreading sequence, τ is the time-delay, $m(t) \in \{-1, 1\}$ is the binary navigation message data, and $n(t)$ is white Gaussian noise with power spectral density $N_0/2$. Thus, the PR sequence can be represented as a convolution

$$c(t) = \sum_{k=-\infty}^{\infty} d_k \sqrt{T_c} p(t - kT_c), \quad (4.36)$$

where $p(t)$ denotes the chip pulse shape which is not necessarily restricted to be time-limited to only one chip interval T_c , that is, the PR sequence $c(t)$ can contain overlapping pulses, so interchip interference can be present. The PR sequence can be assumed as a binary, zero mean wide-sense cyclostationary (WSCS) [4.12, p. 473] sequence with $\{d_k\} \in \{-1, 1\}$ and has period $T = N_d T_c$. $N_d \in \mathbb{N}$ denotes the number of chips of the PR sequence $c(t)$. The multiplication of the navigation data sequence $m(t)$ with the much faster oscillating PR

sequence $c(t)$ introduces a spreading of the spectrum of the navigation signal by a factor of

$$G = \frac{B}{B_m}, \quad (4.37)$$

where B is the one-sided bandwidth of the PR sequence $c(t)$ and B_m denotes the one-sided bandwidth of the navigation message data signal $m(t)$. G is often called spreading gain or processing gain. The signal-to-noise ratio (SNR) after despreading, that is, after correlation of the signal $y(t)$ with $c(t - \tau)$ in the receiver, can be given by

$$\text{SNR}_d = \frac{P}{\tilde{\sigma}_n^2} = \text{SNR}_s G = \frac{P}{\sigma_n^2} G, \quad (4.38)$$

where SNR_s denotes the SNR before despreading, P denotes the signal power, $\sigma_n^2 = BN_0$ is the noise power before despreading, and $\tilde{\sigma}_n^2 = B_m N_0$ is the noise power after despreading. In general $B_m \ll B$. Furthermore, it is assumed that

$$\frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} c(t) c^*(t) dt = 1, \quad (4.39)$$

where the superscript $*$ denotes complex conjugation. For DS-CDMA systems that use a rectangular chip pulse shape $p(t)$ the processing gain can also be given by

$$G = \frac{T}{T_c}. \quad (4.40)$$

The processing gain G and change in bandwidth of a DS-CDMA signal with rectangular chip pulse shape $p(t)$ by despreading is illustrated in Fig. 4.8.

Spread spectrum systems have interference rejection capability (Chap. 16). Let the interference power at the input of the correlator (matched filter) be J , and assume it is uniformly distributed across the spread spectrum bandwidth B (wideband interference). Consequently, we can assume the average interference power spectral density (PSD) to be $J/2B$. Thus, in this case the signal-to-interference-plus-noise-ratio (SINR) before despreading can be given as

$$\text{SINR}_s = \frac{P}{BN_0 + J}, \quad (4.41)$$

and the SINR after despreading for a wideband interference can be given by

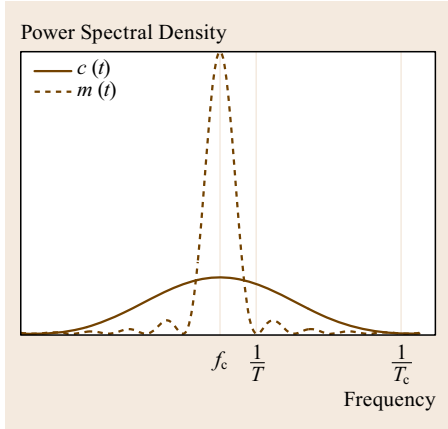


Fig. 4.8 Spread and despread signal with carrier frequency f_c

$$\text{SINR}_d = \frac{P}{B_m N_0 + (B_m/B)J} = \frac{P}{B_m N_0 + J/G} \quad (4.42)$$

We can observe that for broadband interference the SINR increase after despreading is dependent on the processing gain G . Thus, the system design can incorporate certain interference robustness. In the case of narrowband or partial band interference correlation of the received signal with $c(t - \tau)$ spreads the partial band or single-frequency jamming signal so that it appears as wideband Gaussian noise at the output of the correlator.

The autocorrelation of $c(t)$ can be given as

$$\begin{aligned} R_c(\varepsilon) &= \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} c(t) c^*(t + \varepsilon) dt, \\ &= \int_{-\infty}^{\infty} |P(f)|^2 \Phi_d(f) e^{j2\pi f \varepsilon} df, \\ &= \int_{-\infty}^{\infty} |P(f)|^2 e^{j2\pi f \varepsilon} df, \end{aligned} \quad (4.43)$$

where $P(f)$ denotes the Fourier transform of the chip pulse shape $p(t)$ with

$$\int_{-\infty}^{\infty} |P(f)|^2 df = 1. \quad (4.44)$$

This implies that the WSCS sequence $\{d_k\}$ is not only pseudo random but random, and hence the PSD of the sequence $\{d_k\}$ is given as

$$\Phi_d(f) = 1. \quad (4.45)$$

In general, any deterministically generated sequence is not truly random, but the target of sequence design is to achieve sequences that are random as we will see in Sect. 4.2.2. Hence, for well-designed sequences the above given assumption $\Phi_d(f) = 1$ can be justified, and thus the problem of optimizing cross-correlation and autocorrelation properties of PR sequences $c(t)$ can be treated separately as two different problems: on the one hand the problem of optimizing properties of the WSCS sequence $\{d_k\}$ and on the other hand the problem of optimizing properties of the chip pulse shape $p(t)$. This is a convenient assumption not only for GNSS signal design but also for the analysis and understanding of the properties of chip pulse shapes and PR sequences. A proof for (4.43) can be found in [4.12, p. 473].

4.2.2 Pseudo-Random Binary Sequences

As already mentioned earlier, DS-CDMA systems apply spreading or code sequences to separate different channel users (satellites) which are transmitting at the same time within the available transmission bandwidth. These spreading sequences in case of GNSS usually are so-called PR binary sequences. While such deterministically generated sequences can never be truly random, their design ensures well-defined properties associated with randomness. Those kind of sequences are therefore called pseudo random [4.13].

A PR binary sequence is a sequence a_0, a_1, \dots, a_{K-1} with K code bits (also named as *chips*) with $\{a_k\} \in \{-1, 1\}$, where the binary states are represented by -1 and 1 . We can apply the mapping $\tilde{a}_k = (1 - a_k)/2$ with $\{\tilde{a}_k\} \in \{0, 1\}$ and thus the binary states can be represented by 1 and 0 . This mapping preserves the multiplication property on the numbers ± 1 with modulo 2 addition of 0 and 1 . The representation of the binary states by 0 and 1 is useful when discussing properties of sequences in the following. The representation of the binary states by -1 and 1 is used for waveform generation.

The autocorrelation function of the sequence $\{a_k\}$ or $\{\tilde{a}_k\}$ can be given as

$$\begin{aligned} R_a[\ell] &= R_{\tilde{a}}[\ell] \\ &= \sum_{k=0}^{K-1} a_k a_{k+\ell} = \sum_{k=0}^{K-1} (-1)^{\tilde{a}_k + \tilde{a}_{k+\ell}}. \end{aligned} \quad (4.46)$$

Thus, $R_a[0] = R_{\tilde{a}}[0] = K$ when all code bits align. In order to obtain good autocorrelation properties, it is desirable that all other phases ℓ result in very small values of $R_a[\ell]$ and $R_{\tilde{a}}[\ell]$. In a real system, the sequences are finite of length K and thus are repeated after K code bits and thus the signal is periodic. *Solomon Golomb* and

Guang Gong proposed the following three randomness postulates to measure apparent randomness of binary periodic sequences [4.13]:

- In every period, the number of 0's is nearly equal to the number of 1's. (More precisely, the disparity is not to exceed 1; i. e., $|\sum_{k=0}^{K-1} (-1)^{\tilde{a}_k}| \leq 1$)
- In every period, half the runs have length 1, one-fourth have length 2, one-eighth have length 3, and so on, as long as the number of runs so indicated exceeds 1. Moreover, for each of these lengths, there are equally many runs of 0's and 1's. For a binary sequence $\{\tilde{a}_k\}$ with period K , l consecutive 0's (1's) preceded by 1 (or 0) and followed by 1 (or 0) is called a run of 0's (or 1's) of length l .
- The autocorrelation function $R_a[\ell]$ or $R_{\tilde{a}}[\ell]$ is two-valued and it can be given by

$$R_a[\ell] = \begin{cases} K & \text{if } \ell = 0 \pmod{K}, \\ \tilde{K} & \text{if } \ell \neq 0 \pmod{K}, \end{cases} \quad (4.47)$$

where \tilde{K} is a constant. If $\tilde{K} = -1$ for K odd and $\tilde{K} = 0$ for K even, then we say that the sequence has the (ideal) two-level autocorrelation function.

For the special case $K = 2^n - 1$, the resulting sequences are called maximum length sequences or m-sequences. The following properties of m-sequences can be stated [4.13]:

- In every period, 0's occur $2^{n-1} - 1$ (or 2^{n-1}) times and 1's occur 2^{n-1} (or $2^{n-1} - 1$) times. This property is referred to as the balance property [4.13].
- In every period, runs of 0's (or 1's) of length l : $1 \leq l \leq n-2$ occur 2^{n-2-l} times. A run of 0's of length $n-1$ occurs once and a run of 1's of length n occurs once. This is referred to as the run property.
- The autocorrelation function $R_a[\ell]$ or $R_{\tilde{a}}[\ell]$ is two-valued and it can be given by

$$R_a[\ell] = \begin{cases} K & \text{if } \ell = 0 \pmod{K}, \\ -1 & \text{if } \ell \neq 0 \pmod{K}. \end{cases} \quad (4.48)$$

m-Sequences can be generated by linear feedback shift registers (LFSR). A LFSR is a shift register whose input binary state (0 or 1) is a linear function of its previous states. A Q -stage shift register consists of Q consecutive two-state stages (flip-flops) driven by a clock. At each pulse of the clock the state of each stage is shifted to the next stage in line to the right of the register. In order to convert the Q -stage shift register into a sequence generator a feedback loop is incorporated, which calculates a new term for the left-most stage, based on the states of the Q previous states. At the

right-most stage of the register the generated sequence is outputted. A block diagram of a LFSR is depicted in Fig. 4.9. The Boolean feedback function of a LFSR can be expressed as a modulo 2 sum of the Q -stages $z_q \in \{0, 1\}$ element-wise multiplied by feedback coefficients $c_q \in \{0, 1\}$

$$\begin{aligned} f(z_0, \dots, z_q, \dots, z_{Q-1}) \\ = c_0 z_0 \oplus \dots \oplus c_q z_q \oplus \dots \oplus c_{Q-1} z_{Q-1}. \end{aligned} \quad (4.49)$$

The definition of modulo 2 sum (exclusive or) and multiplication are shown in Fig 4.10.

The content of the Q stages of a LFSR (regarded as a binary number or a binary vector of Q bits in length) is called a state of the shift register. $(\tilde{a}_0, \dots, \tilde{a}_q, \dots, \tilde{a}_{Q-1})$ is called the initial state of the shift register which generates the sequence $\{\tilde{a}_k\}$.

Sequences that can be generated based on LFSRs are widely used in many applications, as the sequences can be generated by the LFSR with very low complexity. At the same time, no memory is needed for storing the required sequences. Still today, sequences that can be generated by a LFSR are preferred, as the cost and area demands of memory in chip design are quite high. Nevertheless, the open service (OS) of Galileo (Chap. 9) applies memory codes, which were especially designed with the objective to achieve good autocorrelation and especially good cross-correlation properties.

Besides autocorrelation properties of PR binary sequences also their cross-correlation properties play an important role in terms of system performance of DS-CDMA systems. Especially, the MA interference shall be minimum. As different signals of different satellites are received asynchronously and as their sequences are also modulated by the binary navigation message data $m(t) \in \{-1, 1\}$ (4.35), different cross-correlation properties have to be assessed for good cross-correlation

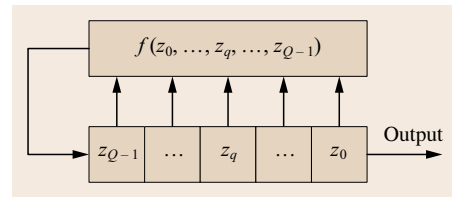


Fig. 4.9 Block diagram of a linear feedback shift register

\oplus	0	1
0	0	1
1	1	0

\cdot	0	1
0	0	0
1	0	1

Fig. 4.10 Definition of modulo-2 sum and multiplication

properties. Only the signals received from one satellite are received synchronously.

As discussed in Sect. 4.2.3, time-delay estimation and subsequently ranging is performed by correlation of the received signal with a local replica of the binary PR sequence used by a certain satellite. Thus, the sequence of the local replica $\{a_k\}$ in general is not synchronous to the sequence $\{b_k\} = b_0, b_1, b_2, \dots, b_K$ with $\{b_k\} \in \{-1, 1\}$ of an interfering signal or of the desired signal $\{a_k\}$ when performing initial correlation in signal acquisition. Furthermore, either the periodic sequence of the interfering signal or the periodic sequence of the desired signal might experience a navigation message data bit transition within the correlation time. So the general situation where the sequence of the local replica $\{a_k\}$ is correlated with part of a sequence of an interfering signal $\{b_k\}$ modulated with a navigation message bit m_0 and with part of a subsequent sequence $\{b_k\}$ modulated with a navigation message bit m_1 needs to be considered.

Thus, we can define the so-called even cross-correlation for $m_1 = 1$ and $m_0 = 1$

$$\begin{aligned} R_{a,b}^e[\ell] &= m_0 \sum_{k=0}^{K-\ell-1} a_k b_{k+\ell} + m_1 \sum_{k=K-\ell}^{K-1} a_k b_{k+\ell} \\ &= \sum_{k=0}^{K-\ell-1} a_k b_{k+\ell} + \sum_{k=K-\ell}^{K-1} a_k b_{k+\ell}, \end{aligned} \quad (4.50)$$

and the so-called odd cross-correlation for $m_0 = 1$ and $m_1 = -1$

$$\begin{aligned} R_{a,b}^o[\ell] &= m_0 \sum_{k=0}^{K-\ell-1} a_k b_{k+\ell} + m_1 \sum_{k=K-\ell}^{K-1} a_k b_{k+\ell} \\ &= \sum_{k=0}^{K-\ell-1} a_k b_{k+\ell} - \sum_{k=K-\ell}^{K-1} a_k b_{k+\ell}. \end{aligned}$$

Both, even and odd cross-correlation are important performance measures in spreading sequence selection or design for DS-CDMA systems and especially for GNSSs, as until today usually no MA interference suppression schemes are applied as discussed in [4.14, 15]. Lloyd Welch [4.16] derived the following bound on even cross-correlation, the so-called Welch's bound. Consider a set of $|S|$ sequences of length K , then for any two sequences $\{a_k\} \in S$ and $\{b_k\} \in S$ with $\{a_k\} \neq \{b_k\}$ the following bound holds

$$\max_{\substack{\{a_k\}, \{b_k\} \in S \\ \{a_k\} \neq \{b_k\}}} \frac{1}{K} R_{a,b}^e[\ell] \geq \sqrt{\frac{|S|-1}{|S|K-1}}. \quad (4.51)$$

Since both $|S|$ and $|S|K$ are typically large, the bound is well approximated by $1/\sqrt{K}$. Welch's bound can be used to derive sequences with good autocorrelation and even cross-correlation properties.

A prominent family of codes that are based on m-sequences and which can be generated by LFSRs are the so-called Gold sequences named after Robert Gold [4.17]. Gold codes are a class of sequences, which provide reasonably large sets of codes with good periodic cross-correlation nearly reaching Welch's bound as well as good autocorrelation properties. They provide better cross-correlation, but worse autocorrelation properties than m-sequences. Gold sequences are produced by modulo 2 sum of two m-sequences each of length $K = 2^n - 1$ in their various phases, and they inherit the balance and run properties of the used m-sequences which have been described before. They have a length of $K = 2^n - 1$ and $|S| = 2^n + 1$ different sequences can be generated. However, if the m-sequences that are modulo 2 added to produce a Gold sequence are chosen just randomly, the cross-correlation of the resulting Gold sequences may be quite poor. Thus, Gold codes are generated using pairs of preferred m-sequences. A method for choosing pairs of preferred m-sequences was given by Robert Gold [4.17]. The preferred sequences $\{g_k\}$ and $\{g'_k\}$ are chosen such that the resulting Gold codes have a three-valued even cross-correlation

$$R_{a,b}^e[\ell] = \begin{cases} -1, \\ -Q, \\ Q-2, \end{cases} \quad (4.52)$$

with

$$Q = \begin{cases} 2^{(n+1)/2} + 1 & \text{if } n \text{ odd,} \\ 2^{(n+2)/2} + 1 & \text{if } n \text{ even.} \end{cases} \quad (4.53)$$

The resulting autocorrelation function is four valued with

$$R_a[\ell] = \begin{cases} -1 & \text{if } \ell \neq 0, \\ -Q & \text{if } \ell \neq 0, \\ Q-2 & \text{if } \ell \neq 0, \\ K & \text{if } \ell = 0. \end{cases} \quad (4.54)$$

Gold codes of the length $K = 1023$ with $n = 10$ are used for the GPS C/A L1 signal (Chap. 7). They are generated with two LFSRs with $Q = 10$ stages. The feedback functions for the two LFSRs are

$$f_1(z_0, \dots, z_9) = z_0 \oplus z_7, \quad (4.55)$$

$$f_2(z_0, \dots, z_9) = z_0 \oplus z_1 \oplus z_2 \oplus z_4 \oplus z_7 \oplus z_8. \quad (4.56)$$

The sequences used for GPS and all other GNSSs, including most Galileo services can be generated by LFSRs, as they are based on m-sequences. They were designed to achieve good autocorrelation and cross-correlation properties as discussed earlier. Galileo E1 OS uses memory codes, which were especially designed with the objective to fulfill the earlier-defined properties for randomness in order to achieve good autocorrelation and cross-correlation properties for application with GNSS (Chap. 9).

For some GNSS signals, additionally so-called tiered codes are applied which are a combination of primary code sequences which are repeated multiplied with code bits of a secondary code. The secondary code is much shorter than the primary code and also repeats periodically. This structure enables to use primary codes only during acquisition in the GNSS receiver while the secondary codes help to further reduce even and odd cross-correlation (Chap. 14) during signal tracking.

4.2.3 Correlation and Time-Delay Estimation

In a GNSS receiver, time-delay estimation is performed using an approximation of a maximum likelihood estimator (MLE). In the receiver, the observations are collected at N time instances with $y[n] = y(nT_s)$, where $T_s = 1/(2B)$ is the sampling duration and $n = 1, 2, \dots, N$. The basic sampling and filtering of the signal using a low-pass filter of bandwidth B with spectral representation

$$H(f) = \begin{cases} 1 & |f| \leq B, \\ 0 & \text{else,} \end{cases} \quad (4.57)$$

and time-domain representation

$$h(t) = 2B \frac{\sin(2\pi Bt)}{2\pi Bt} \quad (4.58)$$

is illustrated in Fig. 4.11. Thus, we can write

$$\mathbf{y} = \sqrt{P}\mathbf{c}(\boldsymbol{\tau}) + \mathbf{n}, \mathbf{y} \in \mathbb{R}^{N \times 1}, \quad (4.59)$$

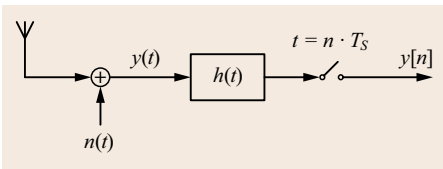


Fig. 4.11 Sampling and low-pass filtering of the received signal

where

$$\begin{aligned} \mathbf{y} &= [y(T_s), \dots, y(nT_s), \dots, y(NT_s)]^\top, \\ \mathbf{n} &= [n(T_s), \dots, n(nT_s), \dots, n(NT_s)]^\top, \\ \mathbf{c}(\boldsymbol{\tau}) &= [c(T_s - \boldsymbol{\tau}), \dots, c(nT_s - \boldsymbol{\tau}), \dots, c(NT_s - \boldsymbol{\tau})]^\top. \end{aligned} \quad (4.60)$$

Let us assume a random variable \mathbf{y} which has a multivariate Gaussian probability density function (pdf) parameterized by the parameter $\boldsymbol{\tau}$, denoted by $p_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\tau})$. Here, \mathbf{y} denotes a realization of \mathbf{y} , which in our case means observations taken in order to estimate $\boldsymbol{\tau}$. After filtering with $h(t)$ the noise is not white Gaussian anymore. Its PSD is given by

$$\Phi_n^B(f) = \begin{cases} \frac{N_0}{2} & |f| \leq B, \\ 0 & \text{else,} \end{cases} \quad (4.61)$$

and its autocorrelation is

$$R_n^B(\varepsilon) = N_0 B \frac{\sin(2\pi Bt)}{2\pi Bt}. \quad (4.62)$$

As $R_n^B(\varepsilon = k/2B) = 0$, with $k \in \mathbb{Z} \setminus \{0\}$ the noise after sampling with $T_s = 1/2B$ is white Gaussian with $\mathcal{N}(0, \sigma_n^2)$ and the noise power is given by $\sigma_n^2 = 2BN_0/2 = BN_0$. Note that, if the sampling duration is chosen $T_s < 1/2B$ or in other words the sampling frequency is chosen $f_s > 2B$, the noise after sampling is not white Gaussian, but so-called colored Gaussian with significant time correlation given by (4.62).

However, if we chose $T_s = 1/2B$ considering the low-pass filter given in (4.58), we can write

$$p_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\tau}) = \frac{1}{(2\pi\sigma_n^2)^{N/2}} \exp\left(-\frac{\|\mathbf{y} - \sqrt{P}\mathbf{c}(\boldsymbol{\tau})\|_2^2}{2\sigma_n^2}\right). \quad (4.63)$$

The likelihood function with respect to the parameter vector $\boldsymbol{\tau}$ is given as

$$L(\mathbf{y}; \boldsymbol{\tau}) = p_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\tau}). \quad (4.64)$$

The likelihood function $L(\mathbf{y}; \boldsymbol{\tau})$ is a function of the parameter $\boldsymbol{\tau}$, which is to be estimated at a given realization of the random variable \mathbf{y} , in our case the vector \mathbf{y} , which contains the samples of the real baseband signal at the output of an antenna. On the other hand, the pdf $p_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\tau})$ is a function of the realization of the random variable \mathbf{y} for a fixed value of the parameter $\boldsymbol{\tau}$. Now the maximum likelihood estimator (MLE) can be given as

$$\begin{aligned} \hat{\boldsymbol{\tau}} &= \arg \max_{\boldsymbol{\tau}} \{L(\mathbf{y}; \boldsymbol{\tau})\} \\ &= \arg \max_{\boldsymbol{\tau}} \{\log(L(\mathbf{y}; \boldsymbol{\tau}))\}. \end{aligned} \quad (4.65)$$

The MLE is asymptotically (large N) unbiased and efficient. When further deriving the estimator we get

$$\begin{aligned}\hat{\tau} &= \arg \max_{\tau} \{\log(L(\mathbf{y}; \tau))\} \\ &= \arg \max_{\tau} \left\{ \log(1) - \log((2\pi\sigma_n^2)^{N/2}) \right. \\ &\quad \left. - \frac{1}{2\sigma_n^2} \|\mathbf{y} - \mathbf{c}(\tau)\|_2^2 \right\} \\ &= \arg \max_{\tau} \left\{ -\|\mathbf{y}\|_2^2 + 2\sqrt{P}\mathbf{y}^\top \mathbf{c}(\tau) \right. \\ &\quad \left. - P\|\mathbf{c}(\tau)\|_2^2 \right\}.\end{aligned}\quad (4.66)$$

As the first term does not depend on τ and the third term is constant with $\|\mathbf{c}(\tau)\|_2^2 \approx N$, $\forall \tau$ as well as dropping the constant factor $2\sqrt{P}$, we can write

$$\hat{\tau} = \arg \max_{\tau} \{\mathbf{y}^\top \mathbf{c}(\tau)\}.\quad (4.67)$$

The MLE of τ is the value for which the power at the output of a correlator matched to the signal is maximized. The cost function is

$$J(\tau) = \mathbf{y}^\top \mathbf{c}(\tau).\quad (4.68)$$

In practice, the maximum of (4.67) is approached using a delay locked loop (DLL), which is the result of applying a gradient ascent method (Chap. 14). Following this gradient ascent method the $k+1$ th iteration can be given as

$$\hat{\tau}^{k+1} = \hat{\tau}^k + \mu \frac{\partial J(\hat{\tau}^k)}{\partial \tau},\quad (4.69)$$

while for the k th iteration the approximation to the maximum is $\hat{\tau}^k$. Here, μ is employed to adjust the step size of the gradient method ($\mu > 0$). In a DLL, the derivative within each iteration is approximated using the central difference quotient of length 2Δ ,

$$\begin{aligned}\hat{\tau}^{k+1} &= \hat{\tau}^k + \frac{\mu}{2\Delta} [J(\hat{\tau}^k + \Delta) - J(\hat{\tau}^k - \Delta)] \\ &= \hat{\tau}^k + \frac{\mu}{2\Delta} \left[\mathbf{y}^\top \mathbf{c}(\hat{\tau}^k + \Delta) - \mathbf{y}^\top \mathbf{c}(\hat{\tau}^k - \Delta) \right].\end{aligned}\quad (4.70)$$

In order to derive the central difference quotient, a common DLL uses a correlator pair where one correlator is advanced by Δ and the other is delayed by Δ , these two correlators are called early and late correlator, respectively. In order to obtain a stochastic version of

the gradient method, we consider that observations are successively collected in intervals with $\mathbf{y}[k]$ being the vector of observations of the k th interval. Thus, we can write

$$\hat{\tau}^{k+1} = \hat{\tau}^k + \frac{\mu}{2\Delta} \left[\mathbf{y}^\top(k+1)\mathbf{c}(\hat{\tau}^k + \Delta) - \mathbf{y}^\top(k+1)\mathbf{c}(\hat{\tau}^k - \Delta) \right].\quad (4.71)$$

The DLL performs the time-delay estimation and provides information for pseudo-range estimates. The DLL is initialized by signal acquisition which provides initial time-delay estimates and initial Doppler frequency estimates (Chap. 14). In a GNSS receiver, the so-called discriminator realizes the earlier-described central difference quotient. Without noise and assuming that the receiver uses signal-matched correlators, the discriminator S-curve for a coherent early-late DLL can be given in terms of the autocorrelation function of the signal $R_c(\varepsilon)$ as

$$S(\varepsilon, \Delta) = R_c(\varepsilon - \Delta) - R_c(\varepsilon + \Delta),\quad (4.72)$$

while the term $\mu/(2\Delta)$ resembles the filter coefficient of the so-called loop filter that reduces the noise, in this case a first-order infinite impulse response (IIR) filter, that drives the DLL (Chap. 14). In (4.72) $\varepsilon = \tau - \hat{\tau}$ denotes the tracking error of the DLL. The basic functionality of a DLL is illustrated in Fig. 4.12.

In order to perform precise synchronization and then positioning the time-delay τ needs to be estimated with high accuracy. If

$$\mathbb{E} \left(\frac{\partial \log(L(\mathbf{y}; \tau))}{\partial \tau} \right) = 0,\quad (4.73)$$

the variance of the time-delay estimation error $\sigma_{\hat{\tau}}^2$ of any unbiased estimator is lower bounded by the Cramer Rao

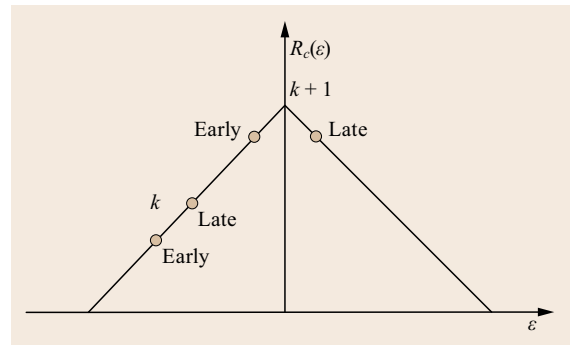


Fig. 4.12 Basic functionality of a DLL

lower bound (CRLB) [4.18]

$$\text{var}(\hat{\tau}) = \sigma_{\hat{\tau}}^2 \geq \frac{1}{-\text{E}\left(\frac{\partial^2 \log(L(y;\tau))}{\partial \tau^2}\right)} = \frac{\frac{\sigma_n^2}{P}}{\frac{\partial \mathbf{c}^\top(\tau)}{\partial \tau} \frac{\partial \mathbf{c}(\tau)}{\partial \tau}},$$

which leads to

$$\sigma_{\hat{\tau}}^2 \geq \frac{B_n}{\frac{P}{N_0} 4\pi^2} \frac{1}{\int_{-\infty}^{\infty} f^2 |P(f)|^2 df}. \quad (4.74)$$

Here, B_n denotes the equivalent noise bandwidth [4.19, 20] of the generic estimator and

$$\int_{-\infty}^{\infty} |P(f)|^2 df = 1. \quad (4.75)$$

A formulation of the CRLB considering quantization effects can be found in [4.21]. The term $\int_{-\infty}^{\infty} f^2 |P(f)|^2 df$ is the second moment of the power spectrum, the so-called **RMS** (root mean square) or Gabor bandwidth [4.22]. This is equal to the curvature of the autocorrelation function $R_c(\varepsilon)$ at $\varepsilon = 0$, as

$$\int_{-\infty}^{\infty} f^2 |P(f)|^2 df = -\frac{1}{4\pi^2} \left. \frac{d^2 R_c(\varepsilon)}{d\varepsilon^2} \right|_{\varepsilon=0} \quad (4.76)$$

from basic theorems of the Fourier transform [4.23, p. 142] and (4.43). Thus, the synchronization accuracy that can be achieved with a GNSS signal in terms of the CRLB increases with the the second moment of the power spectrum of the signal. It can be shown that the second moment of the power spectrum of a signal with bandwidth B is upper bounded by B^2 [4.24]. This means that a higher available signal bandwidth enables a better synchronization accuracy.

In general, higher sidelobes of the autocorrelation function of the signal $R_c(\varepsilon)$ result in higher sidelobes of the chip pulse shape and thus a lower time concentration of the chip pulse shape [4.24, 25]. To quantify the impact of the sidelobe height on the time-delay estimation, we make use of the quantity $\kappa \in [0, 1]$, which denotes the maximum of the absolute values of $R_c(\varepsilon)$ at the local peak values but excluding the the global maximum of at $\varepsilon = 0$.

The higher the value of κ , the less robust the time-delay estimation procedure becomes [4.24, 25]. For the MLE of the time-delay τ , $R_c(\varepsilon)$ can be considered the respective loss function. Time-delay estimation in a nominal GNSS receiver is performed by rather simple methods, for example, an acquisition process or a DLL for time-delay tracking. Hence, it is important for signal design or maybe choice of receiver filtering $h(t)$ to

keep the time sidelobes of $R_c(\varepsilon)$ as small as possible in order to ensure robust time-delay estimation.

As a consequence of the objective of some GNSSs to be interoperable with other GNSSs, they have chosen to use common frequency bandwidths to transmit their signals. Different systems in general use PR binary sequences in order to separate channel users (satellites) from each other. Interoperability with respect to GNSS is defined by the International Committee on Global Navigation Satellite Systems (ICG) as [4.26]:

[...] the ability of global and regional navigation satellite systems and augmentations and the services they provide to be used together to provide better capabilities at the user level than would be achieved by relying solely on the open signals of one system.

Interoperability is often discussed at two different levels, that is, system and signal level. Interoperability at signal level is to be considered in the design of the chip pulse shape as well as the PR binary sequences and is also resulting in MAI-R on top of MAI-A affecting estimation of the time-delay and consequently positioning.

In [4.27], an analysis of MAI-R and MAI-A is presented for different GNSSs. As more and more GNSSs are sharing the same frequency bands, and additionally as they also transmit more and more signals (as well as services) with increasing power, GNSSs become more and more limited not only by noise but also by noise plus MA interference. This may give rise to the application of MA interference suppression methods in the GNSS receiver, as discussed in [4.14, 15].

Following [4.28, pp. 23], [4.29], and other work on chip pulse shape design [4.30–32], MAI-A and MAI-R can be considered as additional interference components with zero mean. In general, both MAI-A and MAI-R are dependent on the propagation characteristics of the transmitted signal. We consider U users (e.g., visible GNSS satellites) with $u = 1, \dots, U$ and power P_u causing MAI-A. Further, we assume that V users of another system (e.g., visible satellites of a different GNSS) with $v = 1, \dots, V$ and power P_v are causing MAI-R. The received signal of another system (e.g., different GNSS) in the same frequency band has PSD $\Phi_R(f)$.

We assume that the reference PR sequence generator is perfectly synchronized with the received desired signal with power P , so the time-delay τ of the desired signal is known. Thus, the receiver was able to perform down conversion, correlation with signal matched reference code, so-called matched filtering with $P^*(f)$, and sampling at the chip duration.

Now, we can define the statistics of the matched filter output for a WSCS sequence $\{d_k\} \in \{-1, 1\}$ with

period $T = N_d T_c$ in the following. The SINR can be given as

$$\text{SINR} = \frac{P}{P_N + P_A + P_R}, \quad (4.77)$$

where the noise power can be given as

$$P_N = \frac{1}{N_d T_c} \frac{N_0}{2} \int_{-\infty}^{\infty} |P(f)|^2 df = \frac{N_0}{2 N_d T_c}, \quad (4.78)$$

the power of the interference component related to MAI-A can be given as

$$P_A = \frac{1}{2 N_d T_c} \sum_{u=1}^U P_u \int_{-\infty}^{\infty} |P(f)|^4 df, \quad (4.79)$$

and the power of the interference component related to MAI-R can be given as

$$P_R = \frac{1}{2 N_d T_c} \sum_{v=1}^V P_v \int_{-\infty}^{\infty} |P(f)|^2 \Phi_R(f) df. \quad (4.80)$$

The background noise is assumed as white Gaussian noise with spectral density of $N_0/2$. The variance due to background noise can be considered as white noise of density $N_0/2$ filtered by the transfer function of the receive filter $P^*(f)$ (correlation with signal matched reference code). Hence, the power of the background white noise can be given by (4.78).

All U users (e.g., visible GNSS satellites) are assumed to be independent and unsynchronized with the desired signal. It is assumed that their time-delays are independently uniformly distributed in $[0, T_c]$ and their phases are independently uniformly distributed in $[0, 2\pi]$. Thus, following [4.28, p. 28], [4.11, p. 772], and [4.29] the effect of the u th user (e.g., GNSS satellite) on the matched filter output of the desired signal will be that of white noise passed through the tandem combination of two filters with the transfer functions $|P(f)|^2$. Thus, the component related to the MAI-A can be defined as given in (4.79). In GNSS, the term $\int_{-\infty}^{\infty} |P(f)|^2 \Phi_R(f) df$ is often called spectral separation coefficient (SSC) and the term $\int_{-\infty}^{\infty} |P(f)|^4 df$ is often called self-SSC [4.29].

Similar assumptions as for the U users above can be taken for the V users and thus the component related to MAI-R can be given as in (4.80). Hence, the ratio of the SNR to the SINR can be given as

$$\Delta \text{SNR} = \frac{\text{SNR}}{\text{SINR}}. \quad (4.81)$$

In order to compare different final pulse shapes $p(t)$ with different values of the design parameter κ and other common chip pulse shapes we adopt the CRLB for time-delay estimation as given in (4.74) and we define the CRLB-I as the CRLB which considers noise plus interference (MAI-A, MAI-R)

$$\tilde{\sigma}_{\hat{\tau}}^2 \geq \sigma_{\hat{\tau}}^2 \Delta \text{SNR}. \quad (4.82)$$

The variance of the MAI-A component as given in (4.79) can be lower bounded by

$$\int_{-B}^B |P(f)|^4 df \geq \frac{1}{2B}, \quad (4.83)$$

subject to

$$\int_{-B}^B |P(f)|^2 df = 1. \quad (4.84)$$

A proof for this lower bound can be found in [4.28, pp. 33].

In order to illustrate the earlier discussed signal or pulse-shape properties we consider three example pulse shapes $p(t)$ with $\kappa = 0.1$, $\kappa = 0.5$, $\kappa = 0.7$, $B = 1.023 \text{ MHz}$, and $BT_c = 1$. Such pulse shapes can be generated following the GNSS signal design methodologies described in [4.24, 25]. The pulse shapes $p(t)$ with $\kappa = 0.1$, $\kappa = 0.5$, and $\kappa = 0.7$, the respective $|P(f)|^2$, and their autocorrelation functions $R_c(\epsilon)$ are depicted in Figs. 4.13, 4.14, and 4.15, respectively.

The CRLB-I for these pulse shapes is shown in Fig. 4.16, where we choose $P_u = -154 \text{ dBW}$ as defined for the Galileo open service (OS) [4.27, 33]

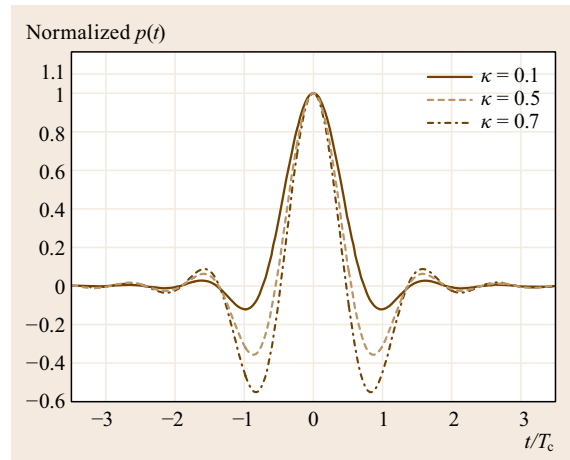


Fig. 4.13 Time domain of example pulse shapes

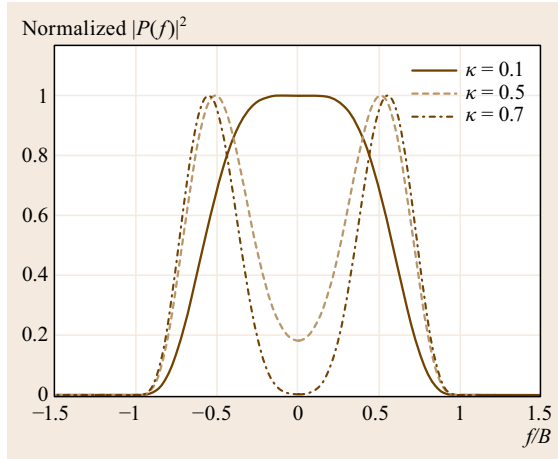


Fig. 4.14 Frequency domain of example pulse shapes

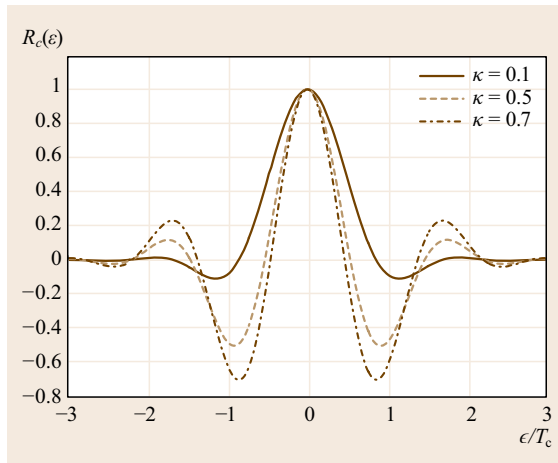


Fig. 4.15 Autocorrelation function of example pulse shapes

and a maximum number of visible Galileo satellites, which contribute to MAI-A of $U = 11$ and $N_0 = -204$ dBW/Hz [4.27, 33]. For this example, we assume that no MAI-R is present. The resulting MAI-A and Δ SNR are given in Table 4.1.

Furthermore, we also assess the multipath performance for these example pulse shapes (Chap. 15). In GNSS signal design multipath performance can be shown by the multipath error envelope [4.34, p. 555] and [4.35]). The multipath error envelope gives the maximum bias of a nominal DLL in case that in addition to the line-of-sight signal a single reflective multipath signal with signal-to-multipath ration of 6 dB is present. The envelope is defined by the cases if the multipath signal has a relative phase of 0 or of π with respect to the line-of-sight (LOS) signal. The multipath error envelope is dependent on the kind of discriminator

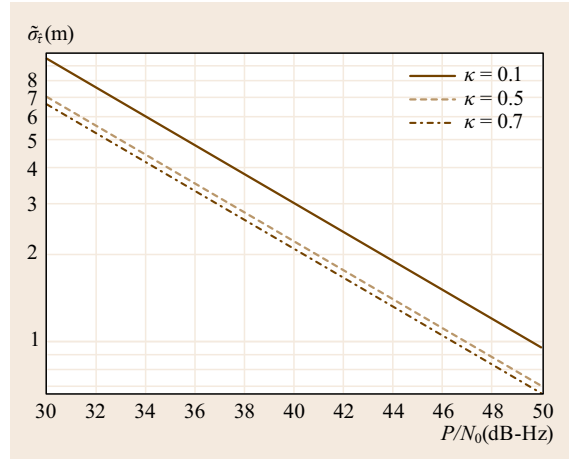


Fig. 4.16 CRLB-I, for example, pulse shapes. The graph shows the time-delay estimation error $\sigma_{\hat{\tau}}$ (4.82) as a function of the power-to-noise-density ratio for a DLL bandwidth of 1 Hz

Table 4.1 Δ SNR and MAI-A, for different example pulse shapes and the lower bound as given in (4.83)

$p(t)$	$\kappa = 0.1$	$\kappa = 0.5$	$\kappa = 0.7$	Lower bound (4.83)
MAI-A	0.79	0.80	0.99	0.54
Δ SNR	1.79 (2.53 dB)	1.80 (2.55 dB)	1.99 (3.0 dB)	1.54 (1.87 dB)

used in the DLL [4.36, 37] and obviously on the shape of the autocorrelation function $R_c(\epsilon)$. The discriminator can be designed in order to optimize tracking performance with respect to a given $R_c(\epsilon)$. We will consider a simple coherent narrow correlator DLL with $2\Delta = 0.1T_c$ two-sided correlator spacing [4.38]. The multipath error envelope for the example pulse shapes with $\kappa = 0.1$, $\kappa = 0.5$, and $\kappa = 0.7$ are provided in Fig. 4.17.

Finally, we also present the loop S-curve of a so-called narrow correlator coherent earlylate DLL in Fig. 4.18. The loop S-curve provides insight on the DLL behavior and also shows the linear region around the main stable lock point. The negative slope of $S(\epsilon, \Delta)$ at $\epsilon = 0$ is equal to the curvature of the autocorrelation function $R_c(\epsilon)$ at $\epsilon = 0$.

Thus, we have assessed two important properties of GNSS signals, time-delay estimation accuracy and on the other hand time concentration and time-delay robustness given by the absolute value of the side-lobes of $R_c(\epsilon)$ given by κ [4.24]. These properties besides others like multipath behavior, MAI-A, MAI-R, signal flexibility, bandwidth efficiency, implementation issues, and others have to be considered when trying to assess ranging performance provided by dif-

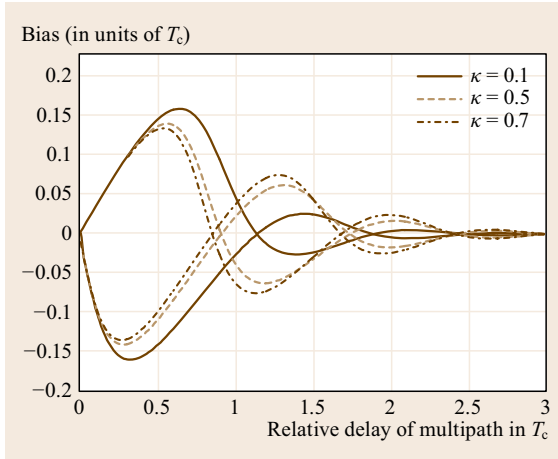


Fig. 4.17 Multipath error envelope for time-delay estimation with example pulse shapes scaled in units of the chip length T_c

ferent GNSS signals. These properties also need to be considered for GNSS signal design and choices of receiver bandwidth B for $h(t)$ of future services and signals.

Let us recapitulate the most important findings on signal properties for spread spectrum ranging from the previous sections:

1. The higher the Gabor bandwidth of the signal, the higher is the synchronization accuracy that can be achieved in terms of the CRLB.
2. Minimizing the CRLB for τ and maximizing time concentration are contradictory tasks.

4.3 Modulation Schemes

In this section, an overview on the most commonly used modulation schemes is provided. The different modulation schemes can also be described by their respective chip pulse shapes. The different common chip pulse shapes for different GNSSs were not only chosen because of the aforementioned properties and their relation to time-delay estimation performance, but also because of an relatively easy implementation on the satellite.

4.3.1 Binary Phase Shift Keying

A rectangular chip pulse shape $p(t)$ can be considered as the *classical* chip pulse shape, which originally was used for early spread spectrum signals [4.11]. If a rectangular chip pulse shape is used the signal is called

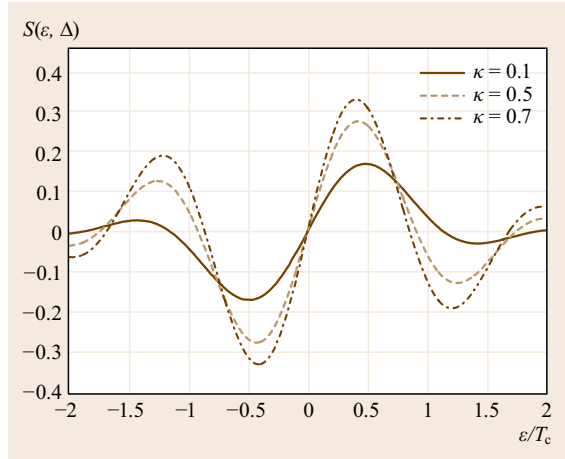


Fig. 4.18 Normalized early-late discriminator S-curve

3. The higher the processing gain G (large bandwidth B), the higher are the synchronization accuracy and the higher interference robustness.
4. The lower the CRLB for τ , the lower is the time concentration of $p(t)$
5. The higher the sidelobes of $R_c(\epsilon)$, the higher are the sidelobes of $p(t)$, the lower is the time concentration of $p(t)$
6. The higher the sidelobes of $R_c(\epsilon)$, the higher κ , the less robust is the estimation of τ (likelihood has local maxima besides the global maximum)
7. MAI-A and MAI-R have to be considered in the signal design (chip pulse-shape design) or can be treated in the receiver (multiuser detection and mitigation).

binary phase shift keying **BPSK** signal. It can be described by

$$p_{\Pi}(t) = \frac{1}{\sqrt{T_c}} \left[U\left(t + \frac{T_c}{2}\right) - U\left(t - \frac{T_c}{2}\right) \right], \quad (4.85)$$

where $U(t)$ denotes the unit step or Heaviside's unit step function

$$U(t) = \begin{cases} 0 & t < 0, \\ 1 & t \geq 0, \end{cases} \quad (4.86)$$

and

$$\int_{-\infty}^{\infty} |p_{\Pi}(t)|^2 dt = 1. \quad (4.87)$$

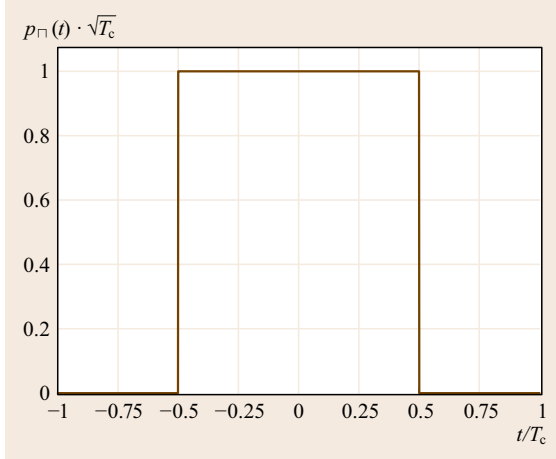


Fig. 4.19 Rectangular chip pulse shape

In Fig. 4.19 $p_Π(t)$ is depicted.

The Fourier transform of $p_Π(t)$ reads

$$\begin{aligned} P_Π(f) &= \frac{1}{\sqrt{T_c}} \left[\frac{1}{2} \delta(f) + \frac{1}{j2\pi f} \right] e^{j2\pi f \frac{T_c}{2}} \\ &\quad - \frac{1}{\sqrt{T_c}} \left[\frac{1}{2} \delta(f) + \frac{1}{j2\pi f} \right] e^{-j2\pi f \frac{T_c}{2}} \\ &= \frac{\sqrt{T_c} \sin(\pi f T_c)}{\pi f T_c} \\ &= \sqrt{T_c} \text{sinc}(f T_c). \end{aligned} \quad (4.88)$$

Here, $\delta(f)$ denotes the Dirac delta function and the sinc function is defined according to [4.23, p. 62]

$$\text{sinc}(t) = \frac{\sin(\pi t)}{\pi t}. \quad (4.89)$$

In Fig. 4.20, the autocorrelation function $R_Π(\varepsilon)$ for a rectangular chip pulse shape is depicted. The autocorrelation function $R_Π(\varepsilon)$ has a triangular shape and can be given as

$$R_Π(\varepsilon) = \begin{cases} 1 - \frac{|\varepsilon|}{T_c} & |\varepsilon| \leq T_c, \\ 0 & \text{else.} \end{cases} \quad (4.90)$$

In Fig. 4.21, the PSD $|P_Π(f)|^2 = T_c \text{sinc}^2(f T_c)$ for a rectangular chip pulse shape is shown. The PSD $|P_Π(f)|^2$ has a main lobe between $-f T_c = -1$ and $f T_c = 1$. Besides the main lobe the PSD has many sidelobes which are decaying rapidly.

As the GNSS signal is band-limited either at the transmitter (satellite payload) or at the receiver, it is useful to also use a formulation for a strictly band-limited

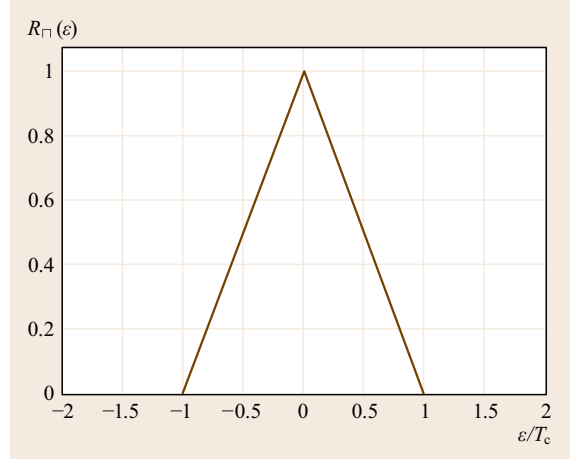


Fig. 4.20 Autocorrelation function for a rectangular chip pulse shape

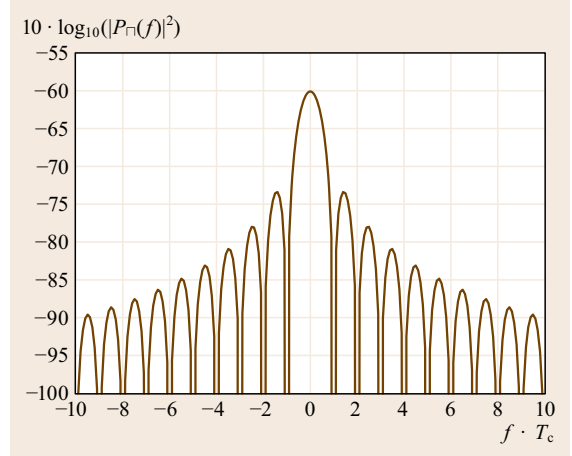


Fig. 4.21 PSD for a rectangular chip pulse shape

and normalized rectangular pulse shape. We assume that the signal is band-limited by an ideal low-pass filter $h(t)$ as defined in (4.58). Such a formulation can be given as

$$p_Π^B(t) = \frac{1}{\xi \pi \sqrt{T_c}} \left(\text{Si} \left[2\pi B \left(t + \frac{T_c}{2} \right) \right] - \text{Si} \left[2\pi B \left(t - \frac{T_c}{2} \right) \right] \right), \quad (4.91)$$

with

$$\text{Si}(t) = \int_0^t \frac{\sin(\tilde{t})}{\tilde{t}} d\tilde{t} \quad (4.92)$$

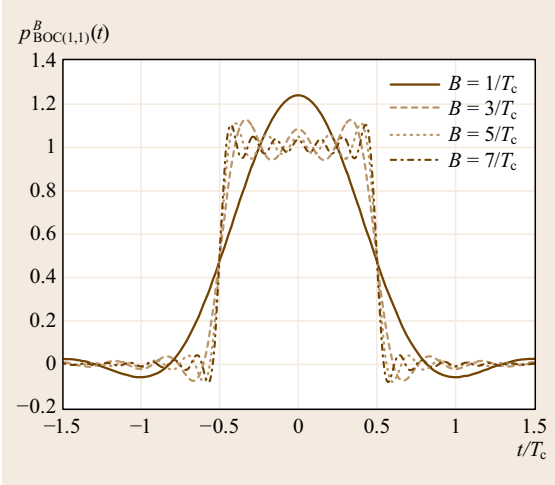


Fig. 4.22 Band-limited and normalized rectangular chip pulse shape

denoting the sine integral function [4.23, p. 62] and with the normalization coefficient

$$\xi = \sqrt{\frac{\int_{-B}^B |P_{\square}(f)|^2 df}{\int_{-\infty}^{\infty} |P_{\square}(f)|^2 df}}. \quad (4.93)$$

In Fig. 4.22 $p_{\square}^B(t)$ is depicted for different bandwidth B in relation to the chip duration T_c . The overshoot or *ringing* of the band-limited rectangular chip pulse shapes is called Gibbs' phenomenon which occurs at the simple discontinuities of the nonband-limited rectangular chip shape when filtering it with an ideal low-pass filter [4.39, p. 30]. We can also observe that due to band-limitation the rectangular chip pulse shape is not time-limited anymore and thus inter-chip interference occurs. The autocorrelation function of the band-limited and normalized rectangular chip pulse shape $R_{\square}^B(\varepsilon)$ is shown in Fig. 4.23.

In comparison to Fig. 4.20, the peak of the autocorrelation function of the band-limited signal $R_{\square}^B(\varepsilon)$ becomes rounded due to the band-limitation of the signal. Thus, the curvature of $R_{\square}^B(\varepsilon)$ at $\varepsilon = 0$ becomes smaller and consequently the CRLB is higher than for the nonband-limited signal.

4.3.2 Binary Offset Carrier Modulation and Derivatives

Binary offset carrier **BOC** signals [4.40, 41] became a kind of standard in GNSS signal design besides using rectangular chip pulse shapes [4.33, 42]. Their chip pulse shapes are formed by the product of a rectangular

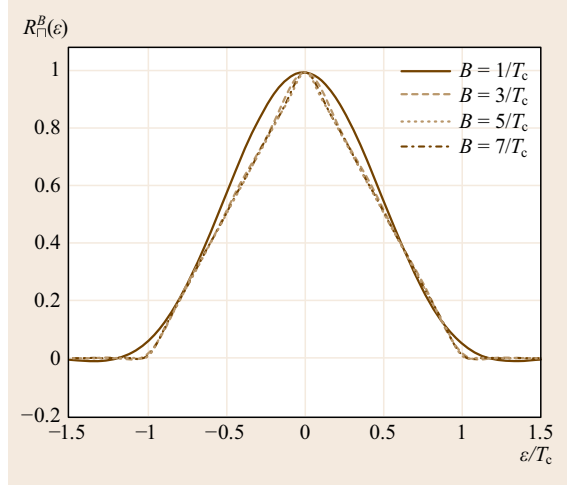


Fig. 4.23 Autocorrelation function of band-limited and normalized rectangular chip pulse shape

pulse

$$p_{nc}(t) = \sqrt{n_c f_r} \left[U\left(t + \frac{1}{2n_c f_r}\right) - U\left(t - \frac{1}{2n_c f_r}\right) \right], \quad (4.94)$$

and a sine or a cosine square wave subcarrier which is given as

$$g_{ns}(t) = \begin{cases} \text{sgn}[-\sin(2\pi n_s f_r t)] \\ \text{sgn}[-\cos(2\pi n_s f_r t)] \end{cases}, \quad (4.95)$$

where n_c and n_s define the chip rate and the subcarrier rate respectively, and f_r is the reference frequency. In general, BOC signals are denoted as $\text{BOC}(n_s, n_c)$ or $\text{BOC}_{\cos}(n_s, n_c)$ for the BOC signals with sine or cosine subcarrier, respectively [4.40, 41] and their pulse shapes are given as

$$p_{\text{BOC}(n_s, n_c)}(t) = \begin{cases} h_{n_c}(t) g_{n_s}(t) & |t| \leq \frac{1}{2n_c f_r} \\ 0 & \text{else} \end{cases}. \quad (4.96)$$

For GNSS signals, for example, for GPS [4.42–44], or the European Galileo system [4.33], $f_r = 1.023$ MHz. The special case of BOC(1,1) with a sine square wave subcarrier is also known as biphase Manchester pulse [4.45, p. 66].

In Figs. 4.24 and 4.25 the chip pulse shapes for a BOC(1,1) signal and a $\text{BOC}_{\cos}(1,1)$ signal are depicted, respectively. When increasing the subcarrier rate several cycles of the binary subcarrier are included within one chip. In the case of $n_s = 4$, we get a chip pulse shape for BOC(4,1) and $\text{BOC}_{\cos}(4,1)$ which com-

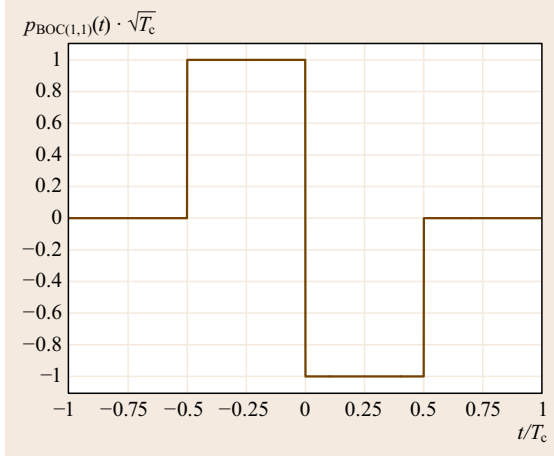


Fig. 4.24 Chip pulse shape for BOC(1,1) signal

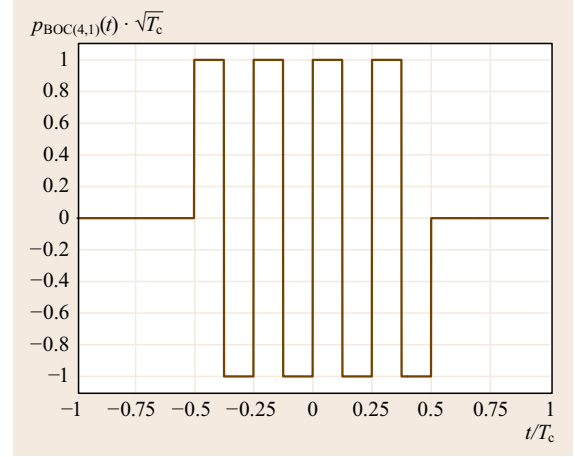


Fig. 4.26 Chip pulse shape for BOC(4,1) signal

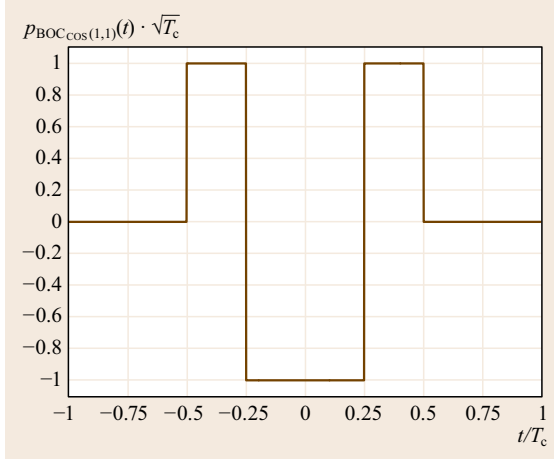


Fig. 4.25 Chip pulse shape for BOC_{cos}(1,1) signal

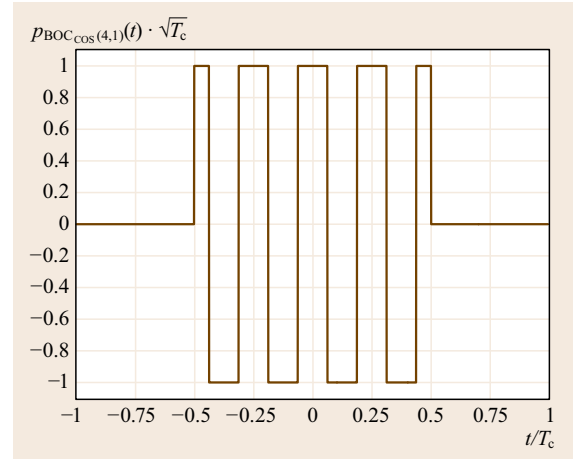


Fig. 4.27 Chip pulse shape for BOC_{cos}(4,1) signal

prise four cycles of the binary subcarrier as shown in Figs. 4.26 and 4.27, respectively.

BOC signals are applied in GNSS in order to fulfill spectral separation requirements between different non-interoperable signals of different GNSS and to enhance synchronization performance [4.33, 40, 42]. Comprehensive derivation of the Fourier transform and the PSD of BOC signals can be found in [4.40, 41, 46]. Thus, following [4.40, 41] we define $n = 2n_s/n_c$ to be the number of half-periods of the subcarrier within the duration of one chip and for BOC signals with sine subcarrier and n even

$$P_{\text{BOC}(n_s, n_c)}(f) = j \sqrt{n_c f_r} \frac{\sin\left(\pi \frac{f}{n_c f_r}\right)}{\pi f} \tan\left(\pi \frac{f}{2n_s f_r}\right), \quad (4.97)$$

for BOC signals with sine subcarrier and n odd

$$P_{\text{BOC}_{\cos}(n_s, n_c)}(f) = \sqrt{n_c f_r} \frac{\cos\left(\pi \frac{f}{n_c f_r}\right)}{\pi f} \tan\left(\pi \frac{f}{2n_s f_r}\right), \quad (4.98)$$

for BOC signals with cosine subcarrier and n even

$$P_{\text{BOC}(n_s, n_c)}(f) = \sqrt{n_c f_r} \frac{\sin\left(\pi \frac{f}{n_c f_r}\right)}{\pi f} \frac{1 - \cos\left(\pi \frac{f}{2n_s f_r}\right)}{\cos\left(\pi \frac{f}{2n_s f_r}\right)}, \quad (4.99)$$

and for BOC signals with cosine subcarrier and n odd

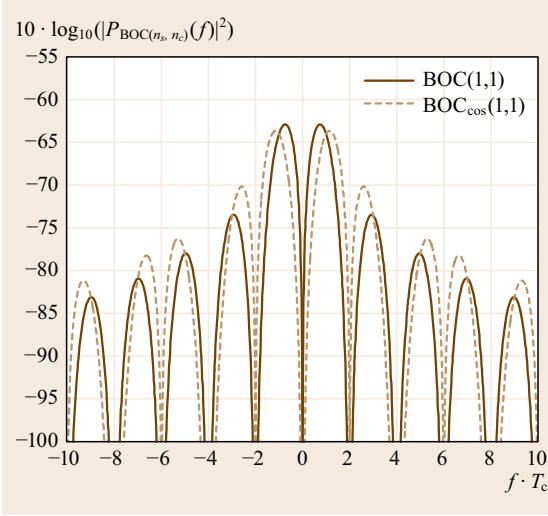


Fig. 4.28 PSD of BOC(1,1) and BOC_{cos}(1,1) signals

$$P_{\text{BOC}_{\cos}(n_s, n_c)}(f) = j \sqrt{n_c f_r} \frac{\cos\left(\pi \frac{f}{n_c f_r}\right) 1 - \cos\left(\pi \frac{f}{2n_s f_r}\right)}{\pi f \cos\left(\pi \frac{f}{2n_s f_r}\right)}, \quad (4.100)$$

where

$$\int_{-\infty}^{\infty} |p_{\text{BOC}(n_s, n_c)}(t)|^2 dt = \int_{-\infty}^{\infty} |P_{\text{BOC}(n_s, n_c)}(f)|^2 df = 1, \quad (4.101)$$

and

$$\int_{-\infty}^{\infty} |p_{\text{BOC}_{\cos}(n_s, n_c)}(t)|^2 dt = \int_{-\infty}^{\infty} |P_{\text{BOC}_{\cos}(n_s, n_c)}(f)|^2 df = 1. \quad (4.102)$$

We can observe that if the chip pulse shape is of even symmetry in time domain then its Fourier transform is real (4.98) and (4.99). On the other hand, in case the chip pulse shape is of odd symmetry in time domain its Fourier transform is imaginary (4.97) and (4.100).

As mentioned earlier, BOC signals are applied in order to fulfill spectral separation requirements between different noninteroperable signals of different GNSS. Following the multiplication of the chips with a sine

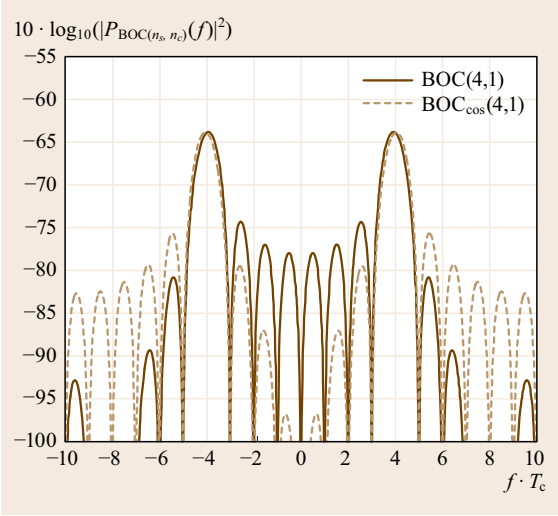


Fig. 4.29 PSD of BOC(4,1) and BOC_{cos}(4,1) signals

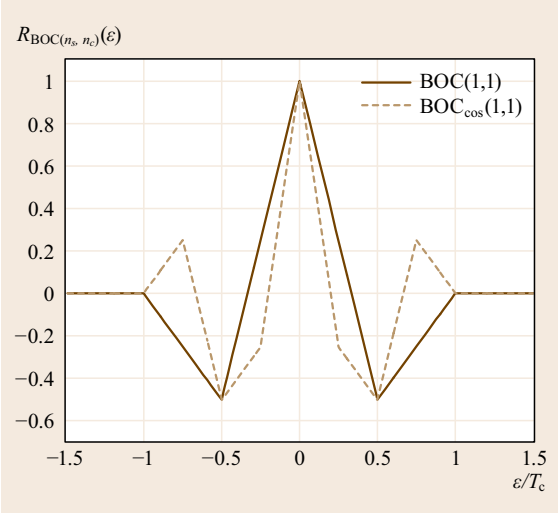


Fig. 4.30 Autocorrelation function of BOC(1,1) and BOC_{cos}(1,1) signals

or cosine binary subcarrier, the spectrum of the signal is divided into two parts; therefore, BOC modulation is also known as a split-spectrum modulation.

In Fig. 4.28, the PSD of a BOC(1,1) and a BOC_{cos}(1,1) signal are shown. The larger the subcarrier rate n_s is chosen, the further the two main lobes of the split-spectrum signal are shifted apart. This effect is illustrated in Fig. 4.29 which shows the PSD of a BOC(4,1) and a BOC_{cos}(4,1) signal.

The autocorrelation function of BOC signals has quite high sidelobes. This is illustrated in Fig. 4.30 for a BOC(1,1) signal and a BOC_{cos}(1,1) signal where the sidelobes' amplitude amounts to $\kappa \approx 0.5$.

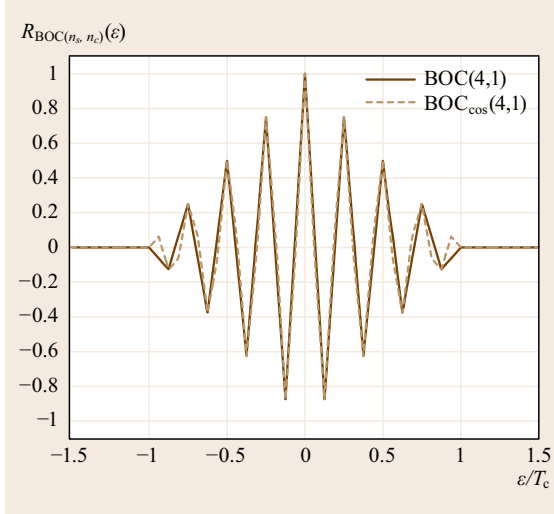


Fig. 4.31 Autocorrelation function of BOC(4,1) and BOC_{cos}(4,1) signals

For larger subcarrier rates n_s the autocorrelation function of split-spectrum signals has even higher sidelobes. In Fig. 4.31, the autocorrelation function of a BOC(4,1) and a BOC_{cos}(4,1) signal are depicted for which $\kappa \approx 0.9$. As discussed in Sect. 4.2.3, high sidelobes of the autocorrelation function, that is, high κ , cause less robust time-delay estimation. Consequently, high κ is to be avoided in order to achieve robust time-delay estimation and signal acquisition.

Further, there exist several extensions of BOC signals, CBOC (composite BOC), MBOC (multiplexed BOC), TMBOC (time-multiplexed BOC), and AltBOC which are applied for the Galileo, GPS, and Beidou signals [4.33, 42, 47]. The AltBOC modulation is rather to be considered a multiplexing/mapping scheme which enables us to multiplex/map several binary signal components to form one signal with a common carrier frequency and thus forming a phase shift keying (PSK) signal. PSK is a digital modulation scheme that conveys data by modulating the phase of a carrier wave. CBOC signals are composed of a linear combination of several BOC signals. A CBOC signal with two BOC signals can be given as

$$p_{\text{CBOC}}(t) = \left[\sqrt{\omega} p_{\text{BOC}(a,b)}(t) \pm \sqrt{1-\omega} p_{\text{BOC}(c,d)}(t) \right], \quad (4.103)$$

where $\omega \in \mathbb{R}_0^+$ and the Fourier transform is given by

$$P_{\text{CBOC}}(f) = \left[\sqrt{\omega} P_{\text{BOC}(a,b)}(f) \pm \sqrt{1-\omega} P_{\text{BOC}(c,d)}(f) \right] \quad (4.104)$$

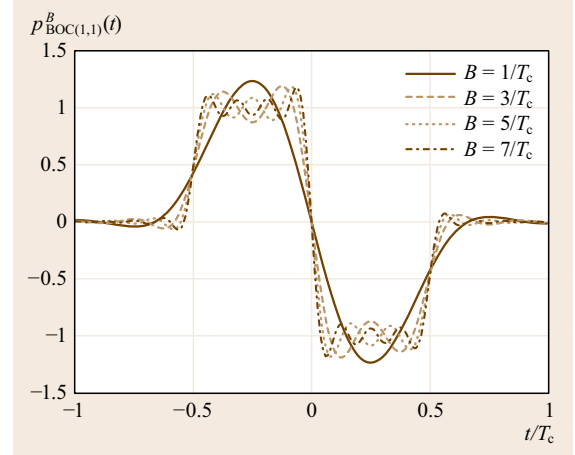


Fig. 4.32 Band-limited BOC(1,1) signal

the PSD is given as

$$|P_{\text{CBOC}}(f)|^2 = \left[\omega |P_{\text{BOC}(a,b)}(f)|^2 + (1-\omega) |P_{\text{BOC}(c,d)}(f)|^2 \pm 2\sqrt{\omega-\omega^2} \text{Re}\{P_{\text{BOC}(a,b)}(f)P_{\text{BOC}(c,d)}^*(f)\} \right]. \quad (4.105)$$

For TMBOC different chip pulse shapes are used for different chips of the PR sequence. A TMBOC chip pulse shape with two different BOC chip pulse shapes which are emitted each T_c seconds can be described by

$$p_{\text{TMBOC}}(t) = \begin{cases} p_{\text{BOC}(a,b)}(t) & \text{probability } p, \\ p_{\text{BOC}(c,d)}(t) & \text{probability } 1-p. \end{cases} \quad (4.106)$$

In case the signaling source, that is, the binary PR sequence with chip pulse shapes, is negative equally probable (NEP), it has the following properties [4.45, p. 64]:

- For each chip pulse shape $p_i(t)$ out of the set of chip pulse shapes that form the signaling source also the negative chip pulse shape $-p_i(t)$ has to be in the set (PR binary sequence $\{d_k\} \in \{-1, 1\}$).
- The stationary probabilities of each chip pulse shape $p_i(t)$ and its negative form $-p_i(t)$ are equal.
- The transition probability $p_{ik} = p_{rs}$ whenever $p_i(t) = \pm p_r(t)$ and $p_k(t) = \pm p_s(t)$, where the transition probability p_{ik} denotes the probability that chip pulse shape $p_k(t)$ is transmitted in the chip after occurrence of the chip pulse shape $p_i(t)$ in the previous chip.

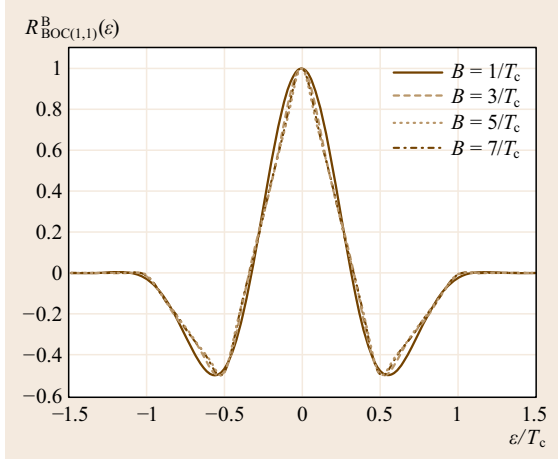


Fig. 4.33 Autocorrelation function of band-limited and normalized BOC(1,1) chip pulse shape

For such a signaling source, the spectrum is characterized by the absence of a line spectrum and further is independent of the transition probabilities themselves. Thus, for a NEP TBOC signal the PSD can be written as

$$|P_{\text{TBOC}}(f)|^2 = p |P_{\text{BOC}(a,b)}(f)|^2 + (1-p) |P_{\text{BOC}(c,d)}(f)|^2, \quad (4.107)$$

and the autocorrelation function is given by

$$R_{\text{TBOC}}(\varepsilon) = p R_{\text{BOC}(a,b)}(\varepsilon) + (1-p) R_{\text{BOC}(c,d)}(\varepsilon). \quad (4.108)$$

4.4 Signal Multiplexing

In the past years, several multiplexing/mapping schemes have been developed and assessed for GNSS use. The overall goal of multiplexing schemes is to map several signals onto one carrier with minimal cross-talk and maximal power and bandwidth efficiency. Moreover, preserving a constant or *quasi*-constant envelope is very beneficial for the amplification of the signals by the high-power amplifier on board the satellite payload, as out-of band emissions and power inefficiencies are mostly avoided. A constant envelope of the signal is given, if the peak-to-average power ratio (PAPR)

$$\text{PAPR} = \frac{\max |x(t)|^2}{E(|x(t)|^2)} = 1, \quad (4.111)$$

where $x(t)$ denotes the multiplexed/mapped signal. Thus, in order to achieve high-power efficiency a small PAPR is desirable.

In the same manner as for the rectangular chip pulse shape, we can also derive strictly band-limited BOC chip pulse shapes. By way of example, we show this in the following for the case of BOC(1,1) with a sine square wave subcarrier, where $T_c = 1/f_r$. The strictly band-limited and normalized BOC(1,1) pulse shape can be given as

$$p_{\text{BOC}(1,1)}^B(t) = \frac{1}{\xi \pi \sqrt{T_c}} \left(2\text{Si}(2\pi B t) \text{Si} \left[2\pi B \left(t + \frac{T_c}{2} \right) \right] - \text{Si} \left[2\pi B \left(t - \frac{T_c}{2} \right) \right] \right), \quad (4.109)$$

with the normalization

$$\xi = \sqrt{\frac{\int_{-B}^B |P_{\text{BOC}(1,1)}(f)|^2 df}{\int_{-\infty}^{\infty} |P_{\text{BOC}(1,1)}(f)|^2 df}}. \quad (4.110)$$

In Fig. 4.32, a band-limited BOC(1,1) chip pulse shape is shown. Similar to the band-limited rectangular pulse in Fig. 4.22 the band-limited BOC(1,1) chip pulse shape is clearly not time-limited anymore. The autocorrelation function of the band-limited and normalized BOC(1,1) chip pulse shape $R_{\text{BOC}(1,1)}^B(\varepsilon)$ is shown in Fig. 4.33. The peak of the autocorrelation function $R_{\text{BOC}(1,1)}^B(\varepsilon)$ becomes rounded due to band-limitation and its curvature is smaller and consequently the CRLB is higher than for the nonband-limited signal.

majority voting [4.51, 52] or [4.53] different modifications like interleave combining [4.52], which can also be considered as a time-code-multiplex approach.

The interplex is particularly interesting because of its power efficiency, when less or equal than five signal components are multiplexed/mapped on one carrier. Moreover, the efficiency is higher if one component is strong and many weak components are combined [4.48]. However, an intermodulation product is needed in order to establish the quasi-constant envelope constellation. We use here the term quasi-constant envelope as in real systems nonbandlimited signals cannot really be generated and thus the signals which are interplexed are not strictly binary anymore. Thus, interplexing/mapping schemes also need to be applicable for nonbandlimited and only quasi-constant envelope signals. In the configuration and power allocation of the current Galileo interplex at E1, around 11% of the total power is spent on the intermodulation product. Even for the very efficient AltBOC modulation for four signal components the power efficiency is around 85%.

In the following, we will study two different multiplexing schemes that are used for GNSS in more detail, the interplex and AltBOC schemes.

4.4.1 Interplex

A phase-modulated radio frequency signal in a phase-shift-keyed/phase modulated (PSK/PM) system can be denoted as [4.48]

$$x(t) = \sqrt{2P} \sin[2\pi f_c t + \Theta(t)], \quad (4.112)$$

where P is the total average power, f_c denotes the carrier frequency, and $\Theta(t)$ is the phase modulation. For an N -channel interplex the phase modulation is

$$\Theta(t) = \left(\beta_1 + \sum_{n=2}^N \beta_n y_n(t) \right) y_1(t), \quad (4.113)$$

where N is the number of channels, β_n are the modulation angles, and $y_n(t) \in \{-1, 1\}$ are the binary data streams or binary GNSS signals.

Two-Channel Interplex with $N = 2$

A two-channel interplex signal is given by

$$x(t) = \sqrt{2P} \sin[2\pi f_c t + \beta_1 y_1(t) + \beta_2 y_1(t) y_2(t)]. \quad (4.114)$$

In the following, we will use the trigonometric identities

$$\begin{aligned} \sin(\alpha \pm \beta) &= \sin(\alpha) \cos(\beta) \pm \cos(\alpha) \sin(\beta), \\ \cos(\alpha \pm \beta) &= \cos(\alpha) \cos(\beta) \mp \sin(\alpha) \sin(\beta). \end{aligned} \quad (4.115)$$

Furthermore, we will use the identities

$$\begin{aligned} \cos[\beta_n y_n(t)] &= \cos(\beta_n), \\ \sin[\beta_n y_n(t)] &= y_n(t) \sin(\beta_n), \end{aligned} \quad (4.116)$$

which are valid for binary signals $y_n(t) \in \{-1, 1\}$. Using these identities, we can reformulate (4.114) to obtain

$$\begin{aligned} x(t) &= \sqrt{2P} \sin(2\pi f_c t) [\cos(\beta_1) \cos(\beta_2) \\ &\quad - y_2(t) \sin(\beta_1) \sin(\beta_2)] \\ &\quad + \sqrt{2P} \cos(2\pi f_c t) [y_1(t) \sin(\beta_1) \cos(\beta_2) \\ &\quad + y_1(t) y_2(t) \cos(\beta_1) \sin(\beta_2)]. \end{aligned} \quad (4.117)$$

Furthermore,

$$\begin{aligned} P_c &= P \cos^2(\beta_1) \cos^2(\beta_2), \\ P_1 &= P \sin^2(\beta_1) \cos^2(\beta_2), \\ P_2 &= P \sin^2(\beta_1) \sin^2(\beta_2), \\ P_{im} &= P \cos^2(\beta_1) \sin^2(\beta_2), \end{aligned} \quad (4.118)$$

are carrier power (P_c), the power in channels 1 and 2 (P_1 , P_2), as well as the power of the intermodulation product (P_{im}).

When choosing $\beta_1 = \pi/2$ and $\beta_2 = \pi/4$, we get

$$\begin{aligned} P_c &= 0, \\ P_2 &= \frac{P}{2}, \\ P_1 &= \frac{P}{2}, \\ P_{im} &= 0, \end{aligned} \quad (4.119)$$

which is equivalent to a quadrature phase shift keying (QPSK) modulation. In Fig. 4.34, such a two-channel interplex signal is shown as equivalent baseband signal

$$\tilde{x}(t) = \sqrt{P_1} y_1(t) + j \sqrt{P_2} y_2(t) \quad (4.120)$$

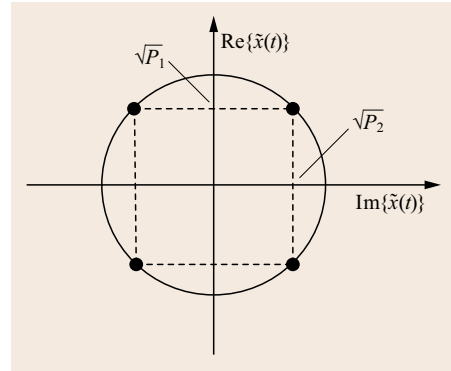


Fig. 4.34 Quadrature phase shift keying (QPSK)

with

$$x(t) = \sqrt{2} \operatorname{Re}\{\tilde{x}(t)e^{j2\pi f_c t}\} \quad (4.121)$$

for the case of equal amplitudes $\sqrt{P_1} = \sqrt{P_2}$. Power balance between real and imaginary part of the QPSK signal is not necessary, though. For GNSS, QPSK is often used with $\sqrt{P_1} \neq \sqrt{P_2}$. The possible phase transitions of standard QPSK are illustrated in Fig. 4.35.

In order to decrease the PAPR (especially when the signals have a finite bandwidth before amplification on the satellite) and thus to avoid phase transitions crossing the origin, staggering of the signal components can be applied. This means that in the case of QPSK the two components are time-delayed with respect to each other by a time-delay τ_s . In general the time-delay τ_s needs to be chosen with respect to the used chip pulse shapes for the signals $y_1(t)$ and $y_2(t)$. A two-channel interplex system with staggering is also called offset-QPSK (OQPSK). The resulting phase transitions for a rectangular chip pulse shape for both $y_1(t)$ and $y_2(t)$ as well as a time-delay $\tau_s = T_c/2$ is depicted in Fig. 4.36. The resulting equivalent baseband signal can be expressed as

$$\tilde{x}_o(t) = \sqrt{P_1}y_1(t) + j\sqrt{P_2}y_2(t - \tau_s). \quad (4.122)$$

Three-Channel Interplex with $N = 3$

A three-channel interplex signal is described by the relation [4.48]

$$x(t) = \sqrt{2P} \sin[2\pi f_c t + \beta_1 y_1(t) + \beta_2 y_1(t)y_2(t) + \beta_3 y_1(t)y_3(t)],$$

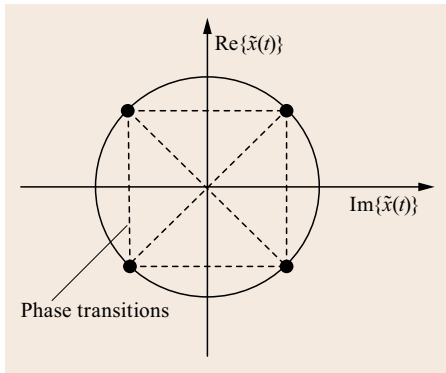


Fig. 4.35 QPSK phase transitions

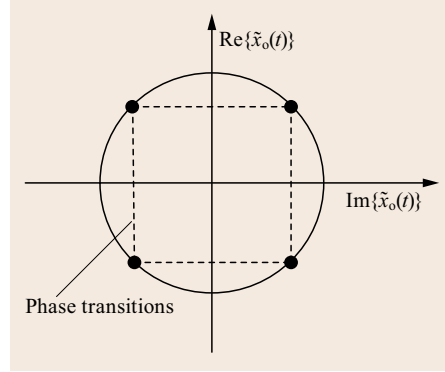


Fig. 4.36 OQPSK phase transitions

which may also be expressed as

$$x(t) = \underbrace{\sqrt{2P} \sin(2\pi f_c t) \cos[\beta_1 y_1(t) + \beta_2 y_1(t)y_2(t) + \beta_3 y_1(t)y_3(t)]}_{=A_1} + \underbrace{\sqrt{2P} \cos(2\pi f_c t) \sin[\beta_1 y_1(t) + \beta_2 y_1(t)y_2(t) + \beta_3 y_1(t)y_3(t)]}_{=A_2}. \quad (4.123)$$

Now, we can write

$$\begin{aligned} A_1 &= \cos(\beta_1) \cos(\beta_2) \cos(\beta_3) \\ &\quad - y_2(t)y_3(t) \cos(\beta_1) \sin(\beta_2) \sin(\beta_3) \\ &\quad - y_2(t) \sin(\beta_1) \sin(\beta_2) \cos(\beta_3) \\ &\quad - y_3(t) \sin(\beta_1) \cos(\beta_2) \sin(\beta_3) \end{aligned} \quad (4.124)$$

and

$$\begin{aligned} A_2 &= y_1(t) \sin(\beta_1) \cos(\beta_2) \cos(\beta_3) \\ &\quad - y_1(t)y_2(t)y_3(t) \sin(\beta_1) \sin(\beta_2) \sin(\beta_3) \\ &\quad + y_1(t)y_2(t) \cos(\beta_1) \sin(\beta_2) \cos(\beta_3) \\ &\quad + y_1(t)y_3(t) \cos(\beta_1) \cos(\beta_2) \sin(\beta_3). \end{aligned} \quad (4.125)$$

In order to eliminate most of the intermodulation terms, we can choose $\beta_1 = \pi/2$ and obtain

$$\begin{aligned} P_1 &= P \cos^2(\beta_2) \cos^2(\beta_3), \\ P_2 &= P \sin^2(\beta_2) \cos^2(\beta_3), \\ P_3 &= P \cos^2(\beta_2) \sin^2(\beta_3), \\ P_{\text{im}} &= P \sin^2(\beta_2) \sin^2(\beta_3), \end{aligned} \quad (4.126)$$

for the powers of the three signal components and the intermodulation product. The equivalent baseband signal for this three-channel interplex signal then can be

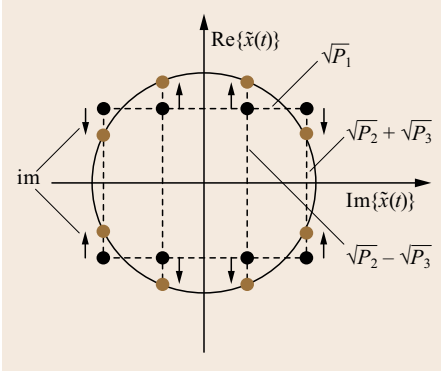


Fig. 4.37 Three-channel interplex signal; impact of intermodulation product (im) shown

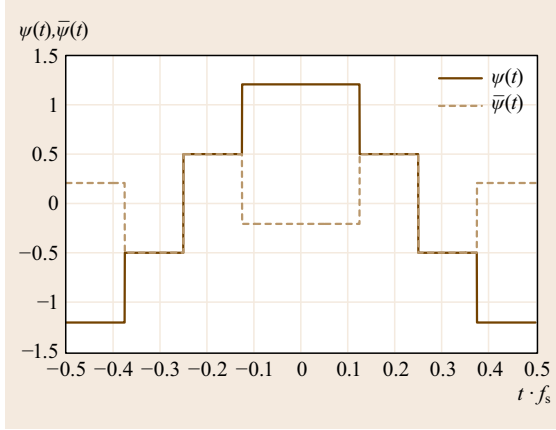


Fig. 4.38 Subcarriers of AltBOC signal

described as

$$\tilde{x}(t) = \sqrt{P_1}y_1(t) - \sqrt{P_{im}}y_1(t)y_2(t)y_3(t) + j \left[\sqrt{P_2}y_2(t) + \sqrt{P_3}y_3(t) \right]. \quad (4.127)$$

This equivalent baseband signal is illustrated in Fig. 4.37 by the brown dots. The black dots denote this interplex signal without any intermodulation product, a nonconstant envelop interplex signal. The effect of the inter-modulation product is indicated by black arrows and the label *im*. Higher orders of interplex signals ($N > 3$) with nearly equal power of the interplex signal components $y_n(t)$ leads to quite high intermodulation power P_{im} and thus high power inefficiency.

In order to increase power efficiency, staggering of all signal components can be applied [4.54] as well as a technique called scalable interplex [4.55] which introduces a shaping of the phase states of the signal constellation (constellation shaping) by weighting different intermodulation products. These techniques provide significant enhancement of power efficiency and they provide the possibility to adapt the interplex signal $x(t)$ to the respective characteristics of the amplifier.

4.4.2 AltBOC

In order to achieve a constant envelope with four signal components, the AltBOC multiplexing/modulation can be used. Four binary signal components can be multiplexed on one carrier frequency with

$$\tilde{x}(t) = \frac{1}{2\sqrt{2}} \left[(y_1(t) + jy_2(t))\psi'_M(t) + (y_3(t) + jy_4(t))\psi_M(t) + (\bar{y}_1(t) + j\bar{y}_2(t))\bar{\psi}'_M(t) + (\bar{y}_3(t) + j\bar{y}_4(t))\bar{\psi}_M(t) \right]$$

with the intermodulation products

$$\begin{aligned} \bar{y}_1(t) &= y_2(t)y_3(t)y_4(t), \\ \bar{y}_2(t) &= y_1(t)y_3(t)y_4(t), \\ \bar{y}_3(t) &= y_1(t)y_2(t)y_4(t), \\ \bar{y}_4(t) &= y_1(t)y_2(t)y_3(t), \end{aligned} \quad (4.128)$$

and the multilevel complex subcarriers

$$\begin{aligned} \psi_M(t) &= \psi(t) + j\psi \left(t - \frac{1}{4f_s} \right), \\ \psi'_M(t) &= \psi(t) - j\psi \left(t - \frac{1}{4f_s} \right), \\ \bar{\psi}_M(t) &= \bar{\psi}(t) + j\bar{\psi} \left(t - \frac{1}{4f_s} \right), \\ \bar{\psi}'_M(t) &= \bar{\psi}(t) - j\bar{\psi} \left(t - \frac{1}{4f_s} \right). \end{aligned} \quad (4.129)$$

The two four-valued subcarriers are given by

$$\begin{aligned} \psi(t) &= +\frac{\sqrt{2}}{4} \text{sgn} \left[\cos \left(2\pi f_s t - \frac{\pi}{4} \right) \right] \\ &\quad + \frac{1}{2} \text{sgn} [\cos(2\pi f_s t)] \\ &\quad + \frac{\sqrt{2}}{4} \text{sgn} \left[\cos \left(2\pi f_s t + \frac{\pi}{4} \right) \right], \end{aligned} \quad (4.130)$$

$$\begin{aligned} \bar{\psi}(t) &= -\frac{\sqrt{2}}{4} \text{sgn} \left[\cos \left(2\pi f_s t - \frac{\pi}{4} \right) \right] \\ &\quad + \frac{1}{2} \text{sgn} [\cos(2\pi f_s t)] \\ &\quad - \frac{\sqrt{2}}{4} \text{sgn} \left[\cos \left(2\pi f_s t + \frac{\pi}{4} \right) \right]. \end{aligned} \quad (4.131)$$

and illustrated in Fig. 4.38. The resulting equivalent baseband signal is shown in Fig. 4.39.

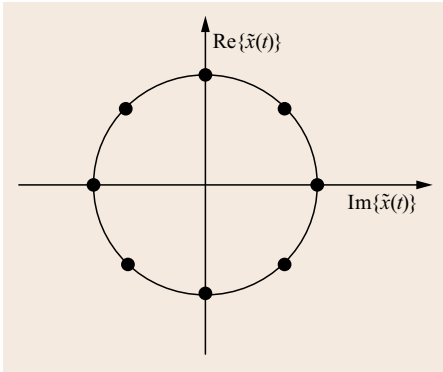


Fig. 4.39 Equivalent baseband signal of AltBOC with $f_s = 15 \times 1.023$ MHz

In the case of four signals with $1/T_c = 10 \times 1.023$ Mcps and rectangular chip pulse shapes as well as a subcarrier frequency $f_s = 15 \times 1.023$ MHz, as used for the Galileo E5 signal we get the spectra for different signals (E5a and E5b) and intermodulation products shown in Fig. 4.40. The sum of all spectra resembles the overall spectrum of the Galileo E5 AltBOC constant envelop signal. Note that after amplification

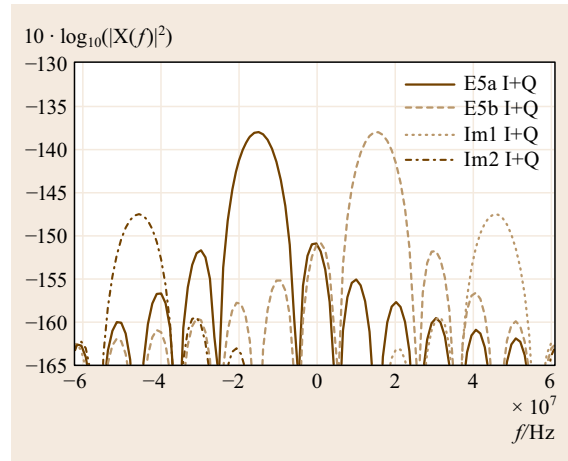


Fig. 4.40 Spectrum of different signal components of AltBOC Galileo E5 signal

on the satellite each payload has also an output filter which filters the AltBOC or interplex signal. Thus, the spectrum of the received signal on ground is then bandlimited with respect to the employed output filter.

4.5 Navigation Data and Data-Free Channels

In general, there are two kinds of channels known as data channel and pilot channels. In data channels, the symbols modulated onto the transmitted PR binary sequences are unknown to the receiver in advance. These symbols are used to transmit the navigation message data $m(t)$ to the user. The navigation message contains all the information that is necessary to allow the user to perform positioning. It includes the ephemeris parameters, needed to compute the satellite coordinates, the time parameters and clock corrections, in order to derive satellite clock offsets and time conversions, several so-called service parameters which indicate satellite health information, parameters to feed the ionospheric model needed for single-frequency receivers, as well as the almanac, which provides orbit and clock information for the entire constellation.

In a pilot channel, the symbols that are modulated onto the transmitted PR binary sequence are known to the receiver. Therefore, a long coherent integration time can be used in the DLL and the code phase estimates can be derived with high accuracy. Also for pilot channels tiered codes are used

in order to reduce cross-correlation, as described in Sect. 4.2.2.

Both data and pilot channel are transmitted synchronously by the same satellite as two different components of the multiplex (interplex, AltBOC, etc.) signal. They use different PR binary sequences in order to be separable by the receiver. Thus, MA interference among data and pilot signals being transmitted synchronously from the same satellite is very low. Data and pilot channels can be processed jointly in the GNSS receiver as they are synchronous and usually coherent.

For some GNSSs, forward error correction (FEC) or channel coding is applied in order to control errors in data transmission noisy over the channel [4.56]. The navigation message data is encoded in a redundant way by using an error-correcting code. The included redundancy allows the GNSS receiver to detect a limited number of errors that may have occurred during symbol detection due to channel impairments and to correct these errors without any retransmission. Error correction and navigation data extraction are further discussed in Sect. 14.5 of this Handbook.

References

- 4.1 J.D. Jackson: *Classical Electrodynamics* (John Wiley, New York 1998)
- 4.2 J.C. Maxwell: *A Treatise on Electricity and Magnetism* (Dover, New York 1979), originally (Oxford Univ. Press 1908)
- 4.3 H. Krim, M. Viberg: Two decades of array signal processing research, *IEEE Signal Process. Mag.* **13**(4), 67–94 (1996)
- 4.4 J. Jeans: *The Mathematical Theory of Electricity and Magnetism* (Cambridge Univ. Press, Cambridge 1908)
- 4.5 J.A. Stratton: *Electromagnetic Theory* (McGraw-Hill, New York 1941)
- 4.6 J.C. Maxwell, P.M. Harman: *The Scientific Letters and Papers of James Clerk Maxwell: 1874–1879* (Cambridge Univ. Press, Cambridge 2002)
- 4.7 D. Zwillinger: *Handbook of Differential Equations* (Academic, San Diego 1997)
- 4.8 B. Hofmann-Wellenhof, H. Lichtenegger, E. Wasle: *GNSS – Global Navigation Satellite Systems – GPS, GLONASS, Galileo and more* (Springer, Vienna 2008)
- 4.9 S. Stein, J.J. Jones: *Modern Communication Principles: With Application to Digital Signaling* (McGraw-Hill, New York 1967)
- 4.10 P. Misra, P. Enge: *Global Positioning System, Signals, Measurements, and Performance* (Ganga-Jamuna, Lincoln 2006)
- 4.11 J.S. Lee, L.E. Miller: *CDMA Systems Engineering Handbook* (Artech House, Norwood 1998)
- 4.12 A. Papoulis, S.U. Pillai: *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York 2002), 4th edn.
- 4.13 S.W. Golomb, G. Gong: *Signal Design for Good Correlation* (Cambridge Univ. Press, Cambridge 2005)
- 4.14 C. Enneking, M. Stein, M. Castaneda, F. Antreich, J.A. Nossek: Multi-satellite time-delay estimation for reliable high-resolution GNSS receivers, *Proc. IEEE/ION PLANS 2012*, Myrtle Beach (ION, Virginia 2012) pp. 488–494
- 4.15 E.P. Glennon, A.G. Dempster: Delayed PIC for post-correlation mitigation of continuous wave and multiple access interference in GPS receivers, *IEEE Trans. Aerosp. Electron. Syst.* **47**(4), 2544–2557 (2011)
- 4.16 L. Welch: Lower bounds on the maximum cross correlation of signals, *IEEE Trans. Inf. Theory* **20**(3), 397–399 (1974)
- 4.17 R. Gold: Optimal binary sequences for spread spectrum multiplexing, *IEEE Trans. Inf. Theory* **13**(4), 619–621 (1967)
- 4.18 S.M. Kay: *Fundamentals of Statistical Signal Processing: Estimation Theory* (Prentice Hall, New Jersey 1993)
- 4.19 R.D. Shelton, A.F. Adkins: Noise bandwidth of common filters, *IEEE Trans. Commun. Technol.* **6**(18), 828–830 (1970)
- 4.20 D.R. White: The noise bandwidth of sampled data systems, *IEEE Trans. Instrum. Meas.* **38**(6), 1036–1043 (1989)
- 4.21 A. Mezghani, F. Antreich, J.A. Nossek: Multiple parameter estimation with quantized channel output, *Proc. Int. ITG/IEEE Workshop Smart Antennas, Bremen* (2010) pp. 143–150
- 4.22 F. Amoroso: The bandwidth of digital data signals, *IEEE Commun. Mag.* **18**(6), 13–24 (1980)
- 4.23 R.N. Barcewell: *The Fourier Transform and its Applications* (McGraw-Hill, New York 1986)
- 4.24 F. Antreich: *Array Processing and Signal Design for Timing Synchronization*, Ph.D. Thesis (Department Electrical Engineering, Munich 2011)
- 4.25 F. Antreich, J.A. Nossek: Optimum chip pulse shape design for timing synchronization, *Proc. IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP, Prague* (2011) pp. 3524–3527
- 4.26 Report of Working Group A: Compatibility and interoperability, ICG/WGA/DEC2008, 3rd Meet. Int. Comm. Glob. Navig. Satell. Syst. (ICG), Pasadena 2008 (2008)
- 4.27 J.V. Perell Gisbert: Interference assessment using up to date public information of operating and under development RNSS systems, Fourth Eur. Work. GNSS Signals Signal Process. (DLR, Oberpfaffenhofen 2009)
- 4.28 A.J. Viterbi: *CDMA: Principles of Spread Spectrum Communication* (Addison-Wesley, Reedwood City 1995)
- 4.29 A coordination methodology for RNSS inter-system interference estimation, Recommendation M.1831-1, Sep. 2015 (ITU, Geneva 2015)
- 4.30 M.A. Landolsi, W.E. Stark: DS-CDMA chip waveform design for minimal interference under bandwidth, phase, and envelope constraints, *IEEE Trans. Commun.* **47**(11), 1737–1746 (1999)
- 4.31 T. Luo, S. Pasupathy, E.S. Sousa: Interference control and chip waveform design in multirate DS-CDMA communication systems, *IEEE Trans. Wirel. Commun.* **1**(1), 56–66 (2002)
- 4.32 M.A. Landolsi: Performance limits in DS-CDMA timing acquisition, *IEEE Trans. Wirel. Commun.* **6**(9), 3248–3255 (2007)
- 4.33 European GNSS (Galileo) Open Service Signal In Space Interface Control Document, OS SIS ICD, Iss. 1.2, Nov. 2015 (EU 2015)
- 4.34 M.S. Braasch: Multipath effects. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker Jr. (AIAA, Washington 1996), pp 547–568, Chap. 14,
- 4.35 M.S. Braasch, A.J. van Dierendonck: GPS receiver architecture and measurements, *Proc. IEEE* **87**(1), 48–64 (1999)
- 4.36 M. Vergara, F. Antreich, G. Artaud, M. Meurer, J.-L. Issler: On performance bounds for GNSS receivers, *Proc. ION GNSS 2009*, Savannah (ION, Virginia 2009) p. 1974
- 4.37 M. Vergara, F. Antreich, M. Meurer: Effect of multipath on code-tracking error jitter of a delay locked loop, *Proc. 4th Eur. Workshop GNSS Signals Signal Process., Oberpfaffenhofen* (2009)
- 4.38 A.J. van Dierendonck, A.J. Fenton: Theory and performance of narrow correlator spacing in a GPS receiver, *Navigation* **39**(3), 265–284 (1992)

- 4.39 A. Papoulis: *The Fourier Integral and its Applications* (McGraw-Hill, New York 1962)
- 4.40 J.W. Betz: Binary offset carrier modulations for radionavigation, *Navigation* **48**(4), 227–246 (2002)
- 4.41 E. Rebeyrol: Galileo Signals and Payload Optimization, Ph.D. Thesis (l'Ecole Supérieure des Télécommunications de Paris, Paris 2007)
- 4.42 Navstar GPS Space Segment/User Segment L1C Interfaces, Interface Specification IS-GPS-800D, 24 Sep. 2013 (Global Positioning Systems Directorate, Los Angeles 2013)
- 4.43 Navstar GPS Space Segment/Navigation User Segment Interfaces, Interface Specification IS-GPS-200H, 24 Sep. 2013 (Global Positioning Systems Directorate, Los Angeles 2013)
- 4.44 Navstar GPS Space Segment/User Segment L5 Interfaces, Interface Specification IS GPS-705D, 24 Sep. 2013 (Global Positioning Systems Directorate, Los Angeles 2013)
- 4.45 M.K. Simon, S.M. Hinedi, W.C. Lindsey: *Digital Communication Techniques, Signal Design and Detection* (Prentice-Hall, New Jersey 1995)
- 4.46 J.A. Avila-Rodriguez: On Generalized Signal Waveforms for Satellite Navigation, Ph.D. Thesis (Department of Aerospace Engineering, University FAF, Munich 2008)
- 4.47 J.-A. Avila-Rodriguez, S. Wallner, G. Hein, E. Rebeyrol, O. Julien, Ch. Macabiau, L. Ries, A. Delatour, L. Lestarcquit, J.-L. Issler: CBOC: An implementation of MBOC, *Proc. 1st CNES-ESA Workshop GALILEO Signals Signal Process.*, Toulouse (2006), hal-01021795
- 4.48 S. Butman, U. Timor: Interplex – An efficient multichannel PSK/PM telemetry system, *IEEE Trans. Commun.* **20**(8), 415–419 (1972)
- 4.49 U. Timor: Equivalence of time-multiplexed and frequency-multiplexed signals in digital communications, *IEEE Trans. Commun.* **20**(8), 435–438 (1972)
- 4.50 P.A. Dafesh, S. Lazar, T. Nguyen: Coherent Adaptive Subcarrier Modulation (CASM) for GPS modernization, *Proc. ION NTM 1999*, San Diego (ION, Virginia 1999) pp. 649–660
- 4.51 G.H. Wang, V.S. Lin, T. Fan, K.P. Maine, P.A. Dafesh: Study of signal combining methodologies for GPS III's flexible navigation payload, *Proc. ION GNSS 2004*, Long Beach (ION, Virginia 2004) pp. 2207–2218
- 4.52 T. Fan, V.S. Lin, G.H. Wang, P.A. Dafesh: Study of signal combining methodologies for future GPS flexible navigation payload (Part II), *Proc. IEEE/ION PLANS 2008*, Monterey (2008) pp. 1079–1108, doi:10.1109/PLANS.2008.4570115
- 4.53 J.J. Spilker Jr., R.S. Orr: Code multiplexing via majority logic for GPS modernization, *Proc. ION GPS 1998*, Nashville (ION, Virginia 1998) pp. 265–273
- 4.54 M. Vergara, F. Antreich: Staggered Interplex, *Proc. IEEE/ION PLANS*, Myrtle Beach 2012 (ION, Virginia 2012) pp. 913–918
- 4.55 M. Vergara, F. Antreich: Evolution of interplex scheme with variable signal constellation, *Proc. ION ITM 2013*, San Diego (ION, Virginia 2013) pp. 651–770
- 4.56 G. Albertazzi, M. Chiani, G.E. Corazza, A. Duverdier, H. Ernst, W. Gappmair, G. Liva, S. Papaharalabos: Forward error correction. In: *Digital Satellite Communications*, ed. by G. Corazza (Springer, New York 2007) pp. 117–174, Chap. 4

Clocks

5. Clocks

Ron Beard, Ken Senior

This chapter provides an overview of clock technology and typical clocks (Cs, Rb, H-Maser) in use today for onboard and ground systems and identifies future trends such as fountain clocks, etc. Concepts such as clock drift, trend, random variations and the statistical methods for their characterization (Allan deviation (ADEV), etc.) are introduced and performance characteristics of global navigation satellite system (GNSS) onboard clocks are presented. The handling and impact of special and general relativity on timing measurements are discussed. Finally, the generation of a GNSS time from an ensemble of ground clocks is described.

5.1	Frequency and Time Stability	122
5.1.1	Concepts	123
5.1.2	Characterization of Clock Stability	123
5.2	Clock Technologies	127
5.2.1	Quartz Crystal Oscillators	127
5.2.2	Conventional Atomic Standards	128
5.2.3	Timescale Atomic Standards	135

5.2.4	Small Atomic Clock Technology	136
5.2.5	Developing Clock Technologies	137
5.3	Space-Qualified Atomic Standards	138
5.3.1	Space Rubidium Atomic Clocks	139
5.3.2	Space-Qualified Cesium Beam Clocks	140
5.3.3	Space-Qualified Hydrogen Maser Clocks ..	141
5.3.4	Space Linear Ion Trap System (LITS)	142
5.3.5	Satellite Onboard Timing Subsystems	142
5.3.6	On-Orbit Performance of Space Atomic Clocks	144
5.4	Relativistic Effects on Clocks	148
5.4.1	Relativistic Terms	148
5.4.2	Coordinate Timescales	150
5.4.3	Geocentric Coordinate Systems	150
5.4.4	Propagation of Signals	153
5.4.5	Relativistic Offset for GNSS Satellite Clocks	154
5.5	International Timescales	155
5.5.1	International Atomic Time (TAI)	155
5.5.2	Coordinated Universal Time (UTC).....	157
5.6	GNSS Timescales	158
	References	160

Today's time and frequency standards range from the most sophisticated reference standards to the smallest oscillators for handheld radios. The technical requirements and technologies needed are different for the various applications but they derive from similar physical concepts. These different technologies can be categorized into four major areas: reference standards, mobile systems, handheld and space systems. These areas are the core areas of time and frequency standard applications and different areas of technology are needed to address them.

Clocks and oscillators are needed by reference timescale centers such as those that contribute to the international time scale, Universal Time Coordinated (UTC). This specialized area requires the most highly stable and accurate time standards that are maintained under controlled environmental conditions. Their outputs are processed with special ensembling algorithms designed to produce an absolute reference for all sys-

tems. For example, the current suite of clocks used at the US Naval Observatory (USNO) consists of many commercial cesium beam frequency standards and hydrogen masers, and specially built rubidium fountain standards. These clocks are physically separated and operated in a tightly controlled environment. Size, weight and power are not issues pertinent for these clocks; primary emphasis is on performance, mostly for intervals of days and much longer.

Clocks used in mobile applications are typically crystal oscillator-based devices and small atomic clocks or oscillators used for positioning, communications or internal to other remote sensing systems. The requirements for mobile devices typically emphasize size, weight and power rather than time and frequency performance so their performance requirements are not particularly demanding or rigorous.

Devices used in handheld applications are the most demanding in terms of size, weight and power. They

commonly use small quartz crystal oscillators. However, in recent years there have been several efforts to develop extremely small atomic standards. These small atom standards offer better accuracy and stability than crystal oscillators in an extremely small package. Although their performance exceeds that of crystal oscillator-based devices they have yet to perform as well as their larger mobile or timing center devices, although there are efforts underway to attempt to improve performance. The technologies developed for these devices have benefited from development in atomic interrogation different from that used by the conventional commercial standards and will be described later.

Space-qualified atomic standards constitute a unique class of frequency standards next to ground- or aircraft-based standards. They were essential for the development and deployment of global navigation satellite systems (GNSS), which are currently the dominant user of highly precise and stable space-qualified

atomic standards. These types of standards provide high stability for navigation performance and a large part of the development of these space devices has been to reliably provide high stability. The GNSS user receiving equipment and the timing capability resulting from their use of the atomic standards in the GNSS satellites produce an inexpensive alternative to high-precision atomic clocks. By displacing higher cost, higher performing atomic clocks, GNSS user equipment receivers or timing receivers with low quality clocks are being deployed in a wide variety of systems. For example, naval tactical and strategic systems are currently utilizing hundreds of Global Positioning System (GPS) units, which are displacing systems on larger ships that may have previously used multiple cesium beam standards on board. Secondary standards such as rubidium vapor cell and crystal oscillators are being used extensively in aircraft, shipboard and man-portable applications, since virtually every system has a clock or oscillator of some quality contained in it.

5.1 Frequency and Time Stability

The oscillator is the basic unit on which clocks, time-keeping and timescales are founded. The fundamental relationship between frequency and time is

$$f = \frac{1}{\tau} \tag{5.1}$$

where f is the frequency of the oscillator, and the period τ is the time interval that a clock uses for time keeping. A clock mechanism added to an oscillator accumulates or counts the number of time intervals to measure elapsed time thereby generating a clock. The intimate connection between oscillators and clocks is sometimes confused by calling a clock a frequency standard or vice versa. A generic clock system is illustrated in Fig. 5.1.

However, oscillators are not perfect and various types have been developed for different requirements

and applications. The determination of oscillator or clock performance under different conditions ranging from ideal laboratory conditions, to harsh field environments, such as military field radios, is an area of special concern.

Oscillators or frequency sources produce noise that appears to be a superposition of causally generated signals and random, nondeterministic noises. Random noises include thermal noise, shot noise, and noises of undetermined origin, such as flicker noise. The end result is time-dependent phase and amplitude fluctuations. Measurement of these fluctuations can be used to characterize the oscillator in terms of amplitude modulation (AM) and phase modulation (PM) noise, and the combination is more commonly called frequency stability. This section describes the basic concepts and measures used to describe the frequency and time stability of precision clocks.

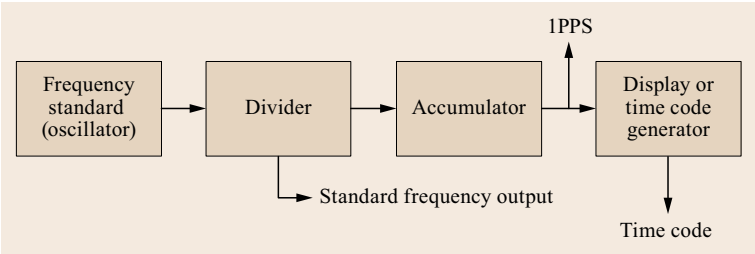


Fig. 5.1 Generic clock system

5.1.1 Concepts

Frequency stability encompasses the concept of random noise, intended and incidental noise, and any other fluctuations in the output frequency of a device. In general, frequency stability is the degree to which an oscillator produces the same frequency throughout a specified period of time. It is implicit in this general definition that the stability of a given frequency decreases if it is anything except a perfect sine wave.

The oscillator produces a signal, whose voltage output may be written as,

$$V(t) = [V_0 + \varepsilon(t)] \sin [2\pi\nu_0 t + \phi(t)] ,$$

where V_0 is the nominal peak voltage amplitude, ν_0 is the nominal fundamental frequency, and where $\varepsilon(t)$ and $\phi(t)$ represent the fluctuations in amplitude and phase of the oscillator from their nominal value, and t represents elapsed time. The instantaneous angular frequency of the oscillator is defined as the time derivative of its total phase

$$\omega(t) = \frac{d}{dt} [2\pi\nu_0 t + \phi(t)]$$

and therefore its instantaneous frequency as $\nu(t) = \omega(t)/(2\pi)$, or

$$\nu(t) = \nu_0 + \frac{1}{2\pi} \frac{d\phi}{dt} . \quad (5.2)$$

For precision oscillators the amplitude fluctuations ε may generally be neglected as they are usually very small compared to the nominal amplitude and therefore have no substantial influence on the frequency or phase. Also, the second term of (5.2) is quite small as compared to the nominal frequency ν_0 and so it is more convenient to define the normalized (or fractional) frequency as

$$y(t) = \frac{\nu(t) - \nu_0}{\nu_0} = \frac{1}{2\pi\nu_0} \frac{d\phi}{dt} , \quad (5.3)$$

which is unitless and which may also be used as a basis for comparing oscillators operating at nominally different frequencies. The phase may then be expressed in units of time as

$$x(t) = \frac{\phi(t)}{2\pi\nu_0} , \quad (5.4)$$

that is,

$$y(t) = x'(t) .$$

A generally applicable model of the time error of a clock, $T(t)$, at an elapsed time, t , after synchronization with another, presumably better clock, can be expressed as

$$T(t) = x_0 + y_0 t + \frac{1}{2} D_0 t^2 + \int_0^t E(t) dt + \varepsilon(t) , \quad (5.5)$$

where x_0 represents the synchronization error or offset at $t = 0$, y_0 represents a constant rate or frequency offset of the clock and D_0 represents a constant frequency drift, and $\varepsilon(t)$ represents all the random deviations of the clock's error. The quantity $E(t)$ represents any remaining systematic nonconstant rate difference due to environmental effects (temperature, radiation, accelerations, etc.).

Although the environmental effects are not usually explicitly modeled such effects can be large and should not be ignored especially in field or operating systems. The random fluctuations are often concentrated upon since they may be measured by statistical means after the systematic components, x_0 , y_0 , and D_0 are removed from the clock data. Characterization of the clock's random error contribution to its total time or frequency error is the subject of the next section on stability.

5.1.2 Characterization of Clock Stability

While no single formal definition of stability exists, the characterization of the stability of a clock may be generally considered as any quantification of the stability of the time and frequency output of the device. A number of different measures of stability have been developed over the years to characterize clocks and numerous papers have been published in more detail than presented here. In particular, the information presented here does not include the various methods and special considerations for measuring clocks and oscillators. For a more comprehensive treatment of the characterization methodology refer to the collection of papers in [5.1], and more recent publications [5.2–5].

In an attempt to make uniform the specifications of clocks through the characterization of their stability the Institute of Electrical and Electronic Engineers (IEEE) in the 1970s made several recommended measures of stability, which can broadly be separated into two analysis areas: sample time-averaged or time domain methods and Fourier spectral or frequency domain methods [5.6]. Both approaches may be related mathematically as shown below, though historically either approach was utilized because the particular methods for measuring clocks' error dictated one method over the other. With the progress in digital processing since

the 1960s both methods can usually be applied for clocks measured today [5.7].

The IEEE has recommended as its primary frequency domain measure of a clock's stability the one-sided spectral density $S_y(f)$ of its fractional frequency, or of its spectral density of phase $S_x(f)$ (or $S_\phi(f)$), which, by properties of derivatives and the Fourier transform, are related by

$$\begin{aligned} S_y(f) &= \left(\frac{f}{\nu_0} \right)^2, \\ S_\phi(f) &= (2\pi f)^2 S_x(f). \end{aligned} \quad (5.6)$$

Here note that f represents the Fourier frequency, which should be distinguished from the frequency output of the clock, ν (or y).

The spectral density $S_x(f)$ may be calculated from observations of its corresponding phase or time signal $x(t)$ by using Fourier transforms. The relationship between the Fourier transform $X(f)$ of the signal and the signal itself is given by

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi j f t} dt. \quad (5.7)$$

However, not all the measurements of $x(t)$ are practically available at all continuous values of t . Given evenly spaced discrete measurements $x(k\tau_0)$ of x , where τ_0 is the smallest sampling interval, and k is an integer, the discrete Fourier transform may be invoked where the integral is replaced by an infinite sum

$$X(f) = \sum_{k=-\infty}^{+\infty} x(k\tau_0) e^{-2\pi j f k \tau_0}. \quad (5.8)$$

If the timing signals are assumed to be periodic, that is $x(t+T) = x(t)$ for all t and some period T , then its Fourier series is also discrete.

Assuming together that the signal x is periodic with period T and there are N evenly spaced phase measurements $x(k\tau_0)$, such that $N\tau_0 = T$, the Fourier series at a finite number of Fourier frequencies may be calculated using just a finite sum

$$X\left(\frac{n}{N\tau_0}\right) = \sum_{k=0}^{N-1} x(k\tau_0) e^{-2\pi j (kn/N)} \quad (5.9)$$

for each $n = 0, 1, \dots, N-1$. In other words, the full Fourier frequency content occurs at an exact finite number of frequencies $f_n = n/(N\tau_0)$ with smallest nonzero Fourier frequency f_1 and largest frequency f_{N-1} . Any

Fourier frequency content occurring at frequencies larger than the Nyquist frequency ($1/(2\tau_0)$, or half the sampling frequency) or content in violation of the periodicity assumption will alias into the spectrum calculated from (5.9).

The spectral density of phase (time) may be calculated from (5.9) by combining the squares of the real and imaginary components and dividing by the total time interval T

$$S_x(f_n) = \frac{\text{Re}\{X(f_n)\}^2 + \text{Im}\{X(f_n)\}^2}{T}. \quad (5.10)$$

A generally applicable model for the random fluctuations of most clocks describes the spectral density of phase as a sum of seven independent pure power laws up to a limiting frequency

$$S_x(f) = \begin{cases} \sum_{\beta=-6}^0 g_\beta f^\beta & \text{for } 0 < f < f_h \\ 0 & \text{for } f_h < f. \end{cases} \quad (5.11)$$

Here, β are integers, the g_β are constants indicating the spectral level of the noise, and f_h is the high frequency cutoff of a low-pass filter. The high frequency cutoff is necessary since variances involving integration of $S_x(f)$ over f would yield unrealistic infinite energies. Also, it is assumed that only integer values of β may be present as the model was mostly empirically derived. For an excellent treatment of the continuous β case see [5.4].

Figure 5.2 shows the phase fluctuations of five simulated clocks having spectral densities of phase, $S_x(f) \sim f^\beta$ for $\beta = 0, -1, -2, -3$, and -4 . $\beta = -5$ and -6 were not included since their variations would outscale the other series in the plot. Note that although the series in the figure show realizations of each power law independently a clock may consist of any or all of the processes simultaneously as per the sum in (5.11). These common clock pure power-law noises are often referred to as white phase noise (WHPH, f^0), flicker phase (FLPH, f^{-1}), random walk phase (RWPH, f^{-2}), flicker frequency (FLFR, f^{-3}), random walk frequency (RWFR, f^{-4}), flicker drift (FLDR, f^{-5}), and random walk drift (RWDR, f^{-6}). It is usually the case that one of the noise processes will dominate at a given Fourier frequency so that a plot of Fourier frequency f versus the logarithm of spectral density of phase $S_x(f)$ would indicate the β noise type that is dominant.

Several time-domain or time-averaged measures have been recommended by the IEEE for characterization of stability. The most well-known measure is the two-sample variance (or Allan variance) for quantify-

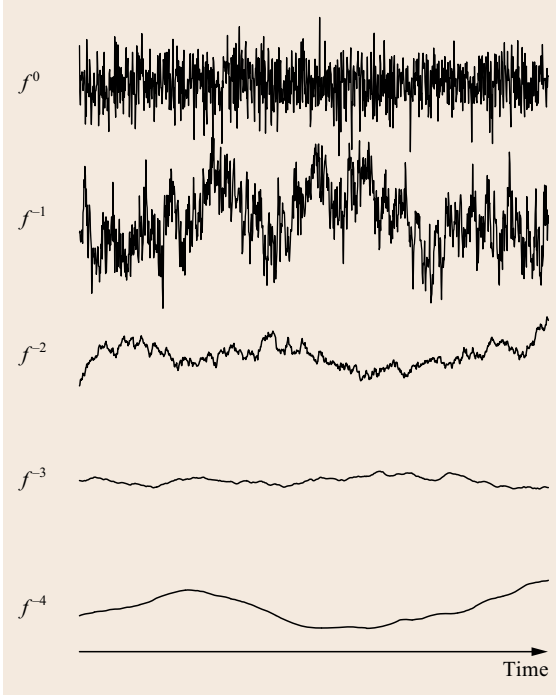


Fig. 5.2 Example simulated realizations of phase (time) fluctuations of five random processes each having respectively from top to bottom spectral densities of phase, $S_x(f) \sim f^\beta$ for $\beta = 0, \dots, -4$

ing frequency stability. It is defined as

$$\sigma_y^2(\tau) = \left\langle \frac{(\bar{y}_{k+1} - \bar{y}_k)^2}{2} \right\rangle, \quad (5.12)$$

where $\langle \cdot \rangle$ denotes infinite time average (or expectation) and where

$$\bar{y}_k = \frac{1}{\tau} \int_{t_k}^{t_k+\tau} y(t) dt = \frac{x_{k+1} - x_k}{\tau}$$

is the average fractional frequency over the interval $\tau = t_{k+1} - t_k$. It is assumed in this definition that the average frequency values are adjacent, that is, no dead time exists between the phase measurement samples x_k . If dead time exists between the samples the resulting calculations will be biased such that the result is no longer considered the Allan variance. The Allan variance is insensitive to an overall systematic frequency or rate offset, y_0 , because fractional frequency averages are differenced in (5.12).

The Allan variance is actually a special case ($N = 2$) of the more general classical N -sample

variance,

$$\sigma_y^2(N, \tau) = \left\langle \frac{1}{N-1} \sum_{i=1}^N \left(\bar{y}_i - \sum_{j=1}^N \bar{y}_j \right)^2 \right\rangle, \quad (5.13)$$

which is an infinite time average of the variance of the N -sample mean of fractional frequency averages. One advantage of the two-sample variance over the N -sample variance is that (5.12) is a well-defined (finite) value for most of the power-law processes in (5.11) (namely for $\beta \geq -4$). Expression (5.13), in contrast, diverges as $N \rightarrow \infty$ for $\beta < 0$ since it depends on the sample interval length T .

Another advantage of the Allan variance is that for power-law processes $\beta = 0, -1, \dots, -4$ in (5.11) the Allan variance $\sigma_y^2(\tau)$ has a τ -relationship similar to the relationship of f to $S_x(f)$. In particular $\sigma_y^2(\tau) \sim |\tau|^\mu$, where $\mu = -3 - \beta$ for $\beta = -2, -3, -4$, while $\mu = -2$ for both $\beta = 0$ and $\beta = -1$. Thus on a log-log plot of τ versus $\sigma_y^2(\tau)$ the slope of the Allan variance curve can be used to indicate the type of noise dominant over that τ interval, except for the WHPH and FLPH noises, which have the same slope τ -Allan variance relationship.

Although the definition of the Allan variance is based on infinite time averaging, a means of estimating it from only a finite portion of the clock's phase realization is required in practice. A common formula used to estimate (5.12) utilizing N discrete samples of the average fractional frequency (or $M = N + 1$ phase samples) is

$$\sigma_y^2(\tau) \approx \frac{1}{2(N-1)} \sum_{i=1}^{N-1} (\bar{y}_{i+1} - \bar{y}_i)^2 \quad (5.14)$$

$$= \frac{1}{2(M-2)\tau^2} \sum_{i=1}^{M-2} (x_{i+2} - 2x_{i+1} + x_i)^2. \quad (5.15)$$

An example clock realization of phase fluctuations is shown in Fig. 5.3, where $N + 1$ measured phase samples are labeled and used to calculate the N average frequencies for a given interval τ . Confidence limits on the variance estimates of (5.14) and (5.12) must also be considered. As in the case of the spectral methods, the estimates obtained are highly dependent on the band-limiting elements of the measurement system. This includes the low-pass filter, which may not be specified as having a sharp cutoff frequency f_h , as assumed in the definition above [5.3].

Confidence limits can be improved by utilizing overlapping samples in the Allan variance calculation. However, the determination of confidence limits is

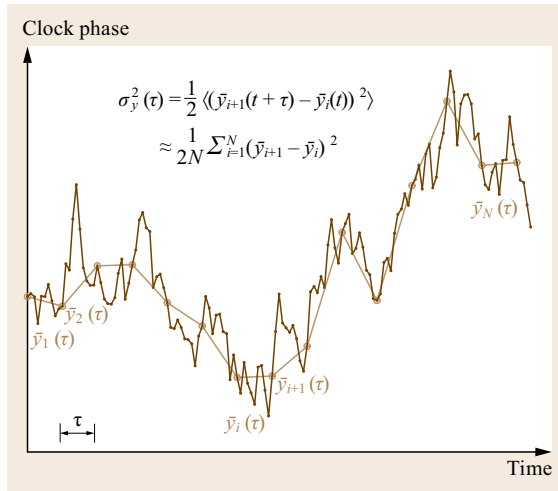


Fig. 5.3 Example of estimating the two-sample (Allan) variance for a given discrete series of clock phase measurements

more complex in this case, since the overlapping fractional frequency samples are no longer independent as in the nonoverlapping case. An estimate of the Allan variance calculated using all possible overlapped \bar{y} samples may be written as

$$\sigma_y^2(\tau) \approx \frac{1}{2m^2(N-2m+1)} \times \sum_{j=1}^{N-2m+1} \left[\sum_{i=j}^{j+m-1} (\bar{y}_{i+1} - \bar{y}_i) \right]^2. \quad (5.16)$$

Most manufacturers of precision clocks or oscillators as well as timing laboratories now routinely publish their clock stability specifications or performances us-

ing the Allan variance (or its square root, the Allan deviation). While the Allan variance may be the most widely used measure of frequency stability there are situations where spectral measures might be preferred. The following formula shows the relationship between the Allan variance and spectral density of phase [5.4]

$$\sigma_y^2(\tau) = \int_0^\infty 2S_x(f) \sin^4(\pi\tau f) df. \quad (5.17)$$

It is valid for all (continuous) $\beta > -5$ and shows that the Allan variance has a very broad Fourier response to energies that are purely harmonic. In this case a spectral approach is preferred over the Allan variance as any such bright lines are more easily identified with frequency domain techniques.

Other time-domain measures of stability include the modified Allan variance, $\text{mod } \sigma_y^2(\tau)$, and the Hadamard variance. The modified Allan variance was introduced in order to address the deficiency of the Allan variance in distinguishing between WHPH and FLPH noises. It does so by effectively varying the (software) bandwidth of the variance calculation to establish the additional τ sensitivity. The Hadamard variance provides yet another time domain measure that converges for all the power-law processes in (5.11) and is insensitive to both an overall systematic rate, y_0 , as well as an overall systematic drift D_0 . Definitions of these variances along with the approximating formulas may be found in [5.1] and [5.5]. Because frequency stability values are commonly expressed as either Allan or Hadamard variances the relationship between the spectral density of phase and these statistics is shown in Table 5.1 for several of the common noise types.

Table 5.1 Relationship between the spectral density of phase $S_x(f)$, the Allan variance $\sigma_y^2(\tau)$, and the Hadamard variance $\text{H}\sigma_y^2(\tau)$ for several common power-law noises

Noise name	Spectral density of phase	Allan variance	Hadamard variance
WHPH	g_0	$\tau^2 \sigma_y^2(\tau) / (3f_h)$	$3\tau^2 \text{H}\sigma_y^2(\tau) / (10f_h)$
RWPH	$g_{-2} f^{-2}$	$\tau \sigma_y^2(\tau)$	$\tau \text{H}\sigma_y^2(\tau)$
RWFR	$g_{-4} f^{-4}$	$3\sigma_y^2(\tau) / \tau$	$6\text{H}\sigma_y^2(\tau) / \tau$
RWDR	$g_{-6} f^{-6}$	$20\sigma_y^2(\tau) / \tau^3$	$120\text{H}\sigma_y^2(\tau) / (11\tau^3)$

5.2 Clock Technologies

As introduced in the previous section, clocks are based on oscillators that generate a periodic signal of a given frequency. The stability of this frequency and the resulting time count depends on the underlying physical principals and design properties and may vary widely between different classes of oscillators.

Key types of oscillators presented in this section include quartz crystal oscillators as well as cesium, rubidium, and hydrogen maser atomic clocks, which constitute the conventional atomic clock technology available today. An overview of the stability that can be expected from the different clock types is shown in Fig. 5.4.

5.2.1 Quartz Crystal Oscillators

The most common and ubiquitous oscillators available are those made with quartz crystals [5.8]. They are a basic form of harmonic oscillator beyond the simple electronic oscillators based on resistor-capacitor (RC) and inductor-capacitor (LC) circuits. Crystal oscillators are used in many forms of electronics and all GNSS receiving equipment operates with these devices to provide the necessary frequencies for radio frequency (RF) signal processing and to form an actual clock.

Quartz is a piezoelectric crystal material that can produce electrical signals by mechanical deformation of the material. Conversely, electrical signals can produce mechanical deformation [5.9]. Crystal oscillators have a higher quality factor (i.e., ratio of resonance

frequency and resonance bandwidth) than the simpler RC and LC circuitry. They have better temperature stability but do use some of the same circuit designs as the LC oscillators with a quartz resonator replacing the tuned circuit portion. Other types of piezoelectric resonators use a surface acoustic waves (SAWs) mechanism, where the signals travel along the surface of the crystal material rather than the more traditional bulk acoustic wave (BAW) mechanism, in which the signals propagate through the crystal. Other types of physically mechanical oscillators are implemented with micro-electro-mechanical system (MEMS) techniques that use devices made from silicon processed through micro-electronic fabrication techniques. The advantage in the MEMS devices is that they are simpler to manufacture and more compatible with modern microelectronic circuitry.

Crystal resonators are available to cover frequencies from about 1 kHz to over 200 MHz. At the low frequency end, wristwatch and real-time clock applications operate at 32.768 kHz and powers of two times this frequency. The conventional BAW resonators range from 80 kHz to 200 MHz. The frequencies of SAW devices range from above 50 MHz to the low GHz range.

The quartz crystal material is comprised of silicon dioxide and can occur naturally or can be grown synthetically. Oscillators are cut from these crystals in a variety of shapes. The shape, size and orientation within the crystalline structure determines the mode of vibration, its resonant frequency and properties of the oscillator. A voltage applied to the crystal will cause it to vibrate and produce a steady signal dependent upon the way the crystal is cut [5.10]. The process of making a quartz oscillator is very involved and complicated requiring material selection, cutting, polishing, mounting electrical contacts and sealing within a vacuum enclosure. Examples of 5 MHz fifth overtone oscillators are shown in Fig. 5.5 without the vacuum enclosure surrounding the crystal. These crystals are mounted on the contacts that extend through the vacuum enclosure to the electrodes plated on the crystal itself.

The quality of a crystal oscillator is determined by its frequency accuracy, frequency stability, aging effects and environmental effects. The absolute frequency accuracy of a crystal oscillator is between 10^{-6} and 10^{-7} taking into account environmental effects such as temperature, mechanical shock and aging. Stability can range from 10^{-10} to 10^{-12} depending upon how protected the crystal is from environmental variations. Aging is defined as the slow change in frequency over a period of time that is associated with long-term changes in the crystal itself or more dominant effects

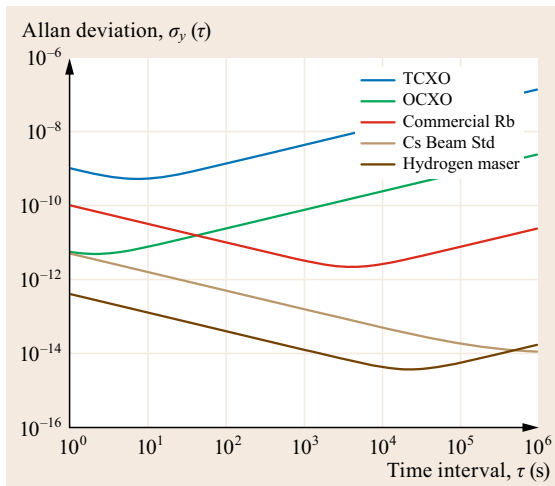


Fig. 5.4 Performance of the classical microwave atomic frequency standards compared with temperature-compensated (TCXOs) and oven-controlled (OCXOs) crystal oscillators

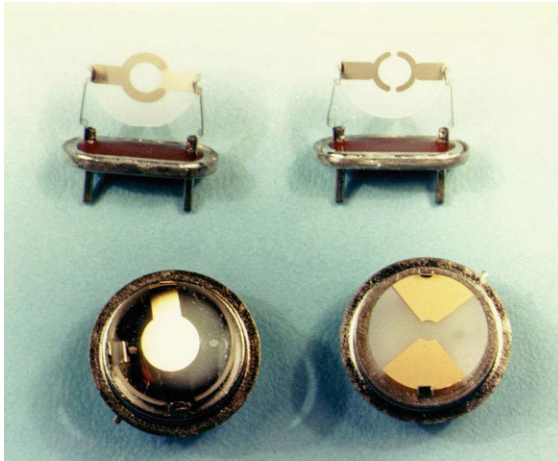


Fig. 5.5 Mounted crystal oscillators with different electrodes. Image courtesy of NRL

such as redistribution of contamination within the vacuum enclosure, slow leaks, mounting and electrode stresses that are relieved with time, and changes in atmospheric pressure. Environmental effects usually have a direct effect on the crystal such as thermal transients, mechanical vibration, shock, radiation, turning the crystal over (tip-over), magnetic fields, voltage changes and variations in the amount of power dissipated in the crystal.

The types of crystal cuts and the method of mitigating the environmental effects on the crystal determines the category of the oscillator. Three configurations in most common use are the room-temperature crystal oscillator (RTXO), the temperature-compensated crystal oscillator (TCXO), and the oven-controlled crystal oscillator (OCXO). The RTXO typically uses a hermetically sealed crystal and individual components for the oscillator circuit. The TCXO encloses the crystal, temperature-compensating components and the oscillator circuit in a container. The OCXO adds heater elements and controls to the oscillator circuit and encloses all the temperature-sensitive components in a thermally insulated container [5.10].

The increased demand for small-scale electronics for cell phones, portable entertainment electronics and miniaturized portable computers has stimulated the development of small-scale quartz oscillators, tuning forks and MEMS oscillators fabricated in silicon. MEMS resonator vibration is based on electrostatic dynamics rather than piezoelectric properties and the MEMS components are micromachined from silicon. They are configured in different complicated shapes such as combs, beam webs, discs and the like that are surrounded by electrodes with transduced gaps on the order of less than 1 μm . All silicon MEMS resonators

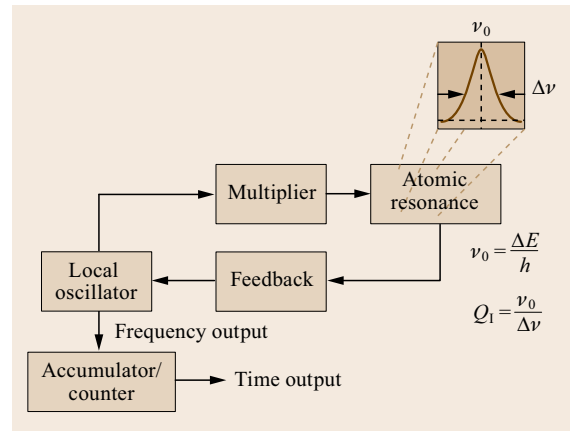


Fig. 5.6 Generic atomic standard block diagram

can be very small and rugged. They are intended for use in integrated circuits at higher frequencies [5.11].

5.2.2 Conventional Atomic Standards

Conventional atomic frequency standard designs are passive devices that are functionally illustrated in Fig. 5.6. The basic principle is to coherently excite transitions between two energy levels in the atom selected and detect that the transition has occurred. The frequency of the atomic transition is

$$\nu = \frac{E_2 - E_1}{h}, \quad (5.18)$$

where E_1 and E_2 are the energy levels of the atom and h is Planck's constant. An important characteristic of the selected transition is the line quality factor

$$Q_1 = \frac{\nu}{\Delta\nu}, \quad (5.19)$$

where $\Delta\nu$ is the line width of the transition. The physics unit generating the precise clock signal incorporates a local oscillator to generate the atomic interrogation signal for the atomic transition and produce the stable output signal locked to the response to that signal. These devices are generally called passive devices because the atomic resonance portion does not actually oscillate but is interrogated with a signal that is modified by the atomic transition to the highly stable, or accurate, signal desired. The interrogation signal is produced by a local oscillator that is typically a quartz crystal oscillator frequency locked to the interrogation signal. The local oscillator may itself be an atomic clock used in a hybrid configuration. Selection or development of the local oscillator can be a significant item in itself.

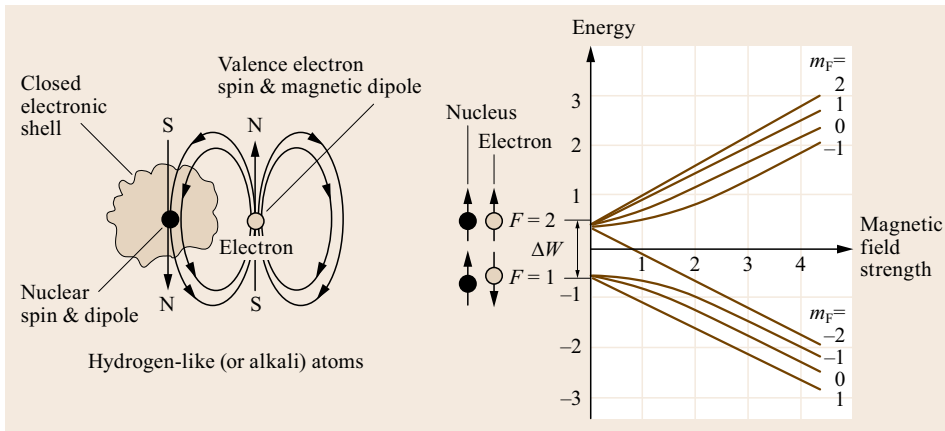


Fig. 5.7 Hyperfine structure of Rb87 with Zeeman splitting

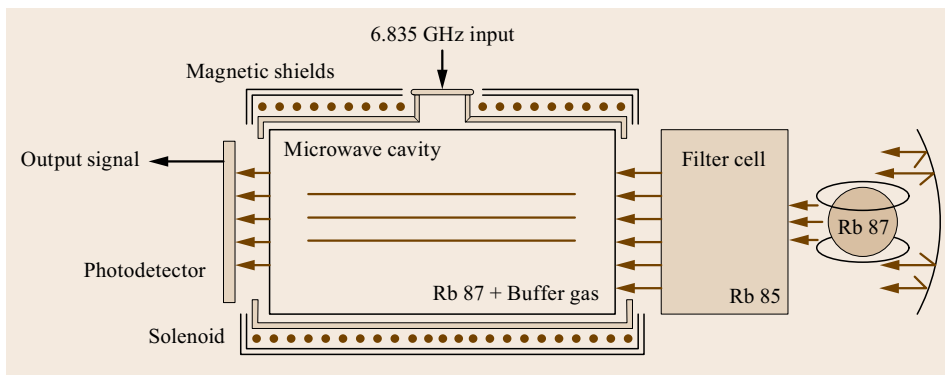


Fig. 5.8 Rubidium gas cell resonator

Rubidium Frequency Standards

Rubidium gas cell standards are the most commonly produced commercial atomic clocks. They are small, consuming relatively low power and are inexpensive in general. They are widely used in the telecommunications industry as frequency references for cellular telephone systems. They are also often found as internal frequency standards in laboratory instrumentation such as frequency counters, signal generators, and signal analyzers. Rubidium clocks were the first atomic clocks used in orbiting spacecraft and have become the primary clock technology used in the GPS satellites.

The rubidium transition used is the hyperfine ground state of Rb87. The hyperfine structure is illustrated in Fig. 5.7. F is the total angular momentum of the Rb87 atom and m_F is the quantized projection of F along a magnetic field. A transition between the two allowed energy states of F (which is reversing the spin of the valence electron) releases or absorbs an energy difference known as the hyperfine frequency of the ground state.

The atomic resonator shown Fig. 5.8 is an optically pumped device consisting of a series of glass cells containing small amounts of rubidium in gaseous

suspension. The state of the rubidium atoms in the resonance cell are selected using light from an Rb87 lamp, which is filtered through a cell containing Rb85. The resonance cell also contains a buffer gas, typically nitrogen and argon or xenon, to hold the rubidium in suspension and to minimize interactions of the rubidium with the cell walls. The lamp is excited to a plasma with radio frequency (RF) energy creating the full set of Rb87 spectral lines. Only one of these lines is desired for interrogating the Rb87 atoms in the resonance cell. The Rb85 filter cell eliminates most of the unwanted spectral light, allowing a higher signal-to-noise ratio (SNR) at the photodetector. The Rb87 atoms in the resonance cell are at a controlled temperature and magnetic field to minimize environmental effects. The resonance cell is contained in a microwave cavity that creates a uniform RF field in the cell. The nominal frequency of the microwave cavity is about 6.834682611 GHz [5.12].

The overall design of a typical rubidium clock is shown in Fig. 5.9. When the 6.834682611 GHz signal is applied to the microwave cavity in the atomic resonator the level of the spectral light transmitted through the resonance cell is affected depending on how close the signal is to the inherent resonance of the Rb87 atoms.

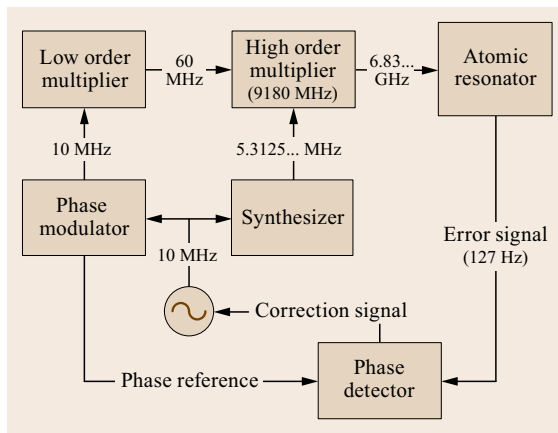


Fig. 5.9 Generic rubidium standard block diagram

An on-resonance signal will cause a decrease in the level of light transmitted through the resonance cell due to the absorption of the spectra. The microwave cavity signal is then modulated at about 127 Hz about the resonance so that the output of the photodetector provides an output signal proportional to the amount of light reaching it. This output signal is used as an error signal for the feedback loop, which adjusts the frequency of the crystal oscillator to minimize the error. The actual output signal is produced by the local oscillator.

This atomic interrogation technique is known as intensity optical pumping (IOP). Another optical technique employing lasers has been developed, known as coherent population trapping (CPT), which has been applied to Rb and Cs gas cell oscillators in smaller physical packages resulting in so-called miniature atomic clocks. This technique and its application to miniature clocks will be discussed later (Sect. 5.2.4).

Rubidium clocks based on the classic IOP design are considered to be secondary frequency standards because their inherent accuracy is significantly affected by the environment and the nature of the gas cell. These effects lead to environmental sensitivity and frequency drift. While rubidium clocks can be set on frequency very precisely, frequency drift rates exceeding 10^{-11} /month reduce absolute accuracy to some parts in 10^{-9} . Gas cell clocks of similar design can also be made using cesium or conceivably other alkali metals. However, using other atoms does not change the basic nature of the clock and does not make them primary standards for the reasons discussed in the next section.

Cesium Beam Frequency Standards

Cesium beam frequency standards are commercially available clocks and have been widely used for time-keeping and precise frequency generation, particularly in the telecommunications industry where they are used

for clocking high-rate data streams. They are inherently much more accurate in frequency than rubidium clocks with accuracies as good as $5 \cdot 10^{-13}$. They also have an inherently very low frequency drift and reduced sensitivity to environmental effects, although the associated electronics in the units may be somewhat affected by environmental conditions, primarily temperature. Specially built cesium beam clocks with large long tubes designed for high accuracy have also been used as primary laboratory standards. Considering the small frequency shifts and the accuracy that can be maintained by a cesium beam frequency standard it is the most accurate device that is easily and commercially available.

The hyperfine frequency of the cesium atom in its ground state is the atomic transition used by these cesium standards. The ground state of Cs133 is the interaction between the hyperfine $F = 3$ and $F = 4$ energy levels. When a magnetic field is applied, the energy levels are divided into sublevels identified by their magnetic quantum number m_F . The frequency of the interaction $F = 3, m_F = 0$ to $F = 4, m_F = 0$ is the Cs hyperfine frequency, $\nu_{\text{hf}} = 9.192631770 \text{ MHz}$.

Because of the increased accuracy available from widely available commercial devices and the primary laboratory standards built for increased accuracy, the atomic second was adopted in 1967 as the basis of time in the Système International (SI) [5.13]. In continuity with previous timescales, the SI second has been defined as [5.14]:

[...] the duration of 9192631770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium 133 atom.

Equivalently, the hyperfine frequency of the cesium atom's ground state amounts to exactly 9 192 631 770 Hz.

The commercial cesium beam tube, Fig. 5.10, is an atomic thermal beam device [5.15, 16]. The frequency ν of the hyperfine transition has a second-order dependence on the applied magnetic field (B in Teslas) of

$$\nu = \nu_{\text{hf}} + 4.27 \cdot 10^{10} \text{ Hz} \cdot B^2.$$

In operation, a cesium reservoir at one end of the sealed vacuum enclosure is heated to about 100°C to produce a small stream of cesium atoms that is collimated into a beam. To limit the atoms in the beam to the useful energy levels, the desired energy states are magnetically selected and atoms in an undesired state are deflected out of the beam. The remaining atoms pass through a two-armed interrogation cavity known as a Ramsey cavity, where they are exposed twice to a microwave field. Here, the atoms change the ground state if the

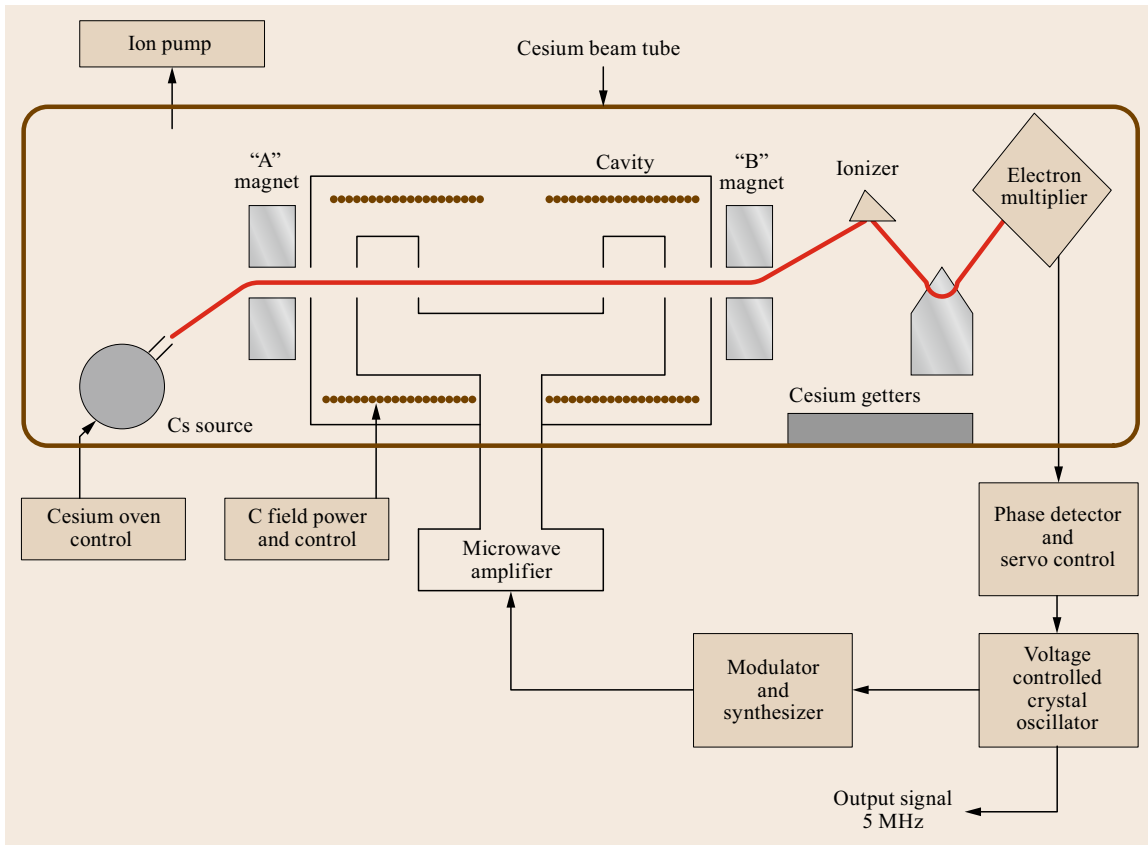


Fig. 5.10 Cesium beam tube diagram

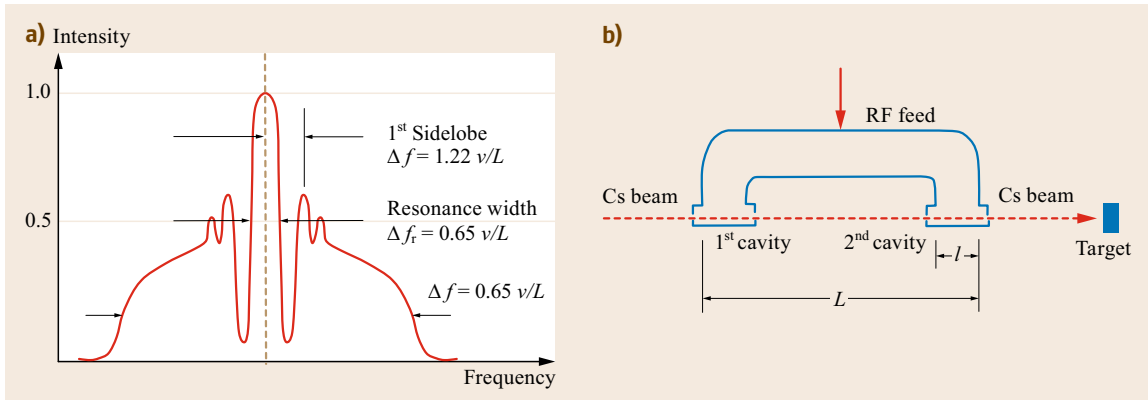


Fig. 5.11a,b Relation between the Ramsey pattern (a) and the dimensions L and l of the cavity (b). v denotes the velocity of atoms in the cesium beam

probing frequency matches the Cs hyperfine frequency. After leaving the cavity, the beam of cesium atoms passes through another magnetic state selector where atoms in the desired state are routed to a detector. Beam current from the electron multiplier is maximized when exactly the right microwave signal is present.

When varying the probing frequency around the nominal value, a resonance pattern with a line width inversely proportional to the cavity length is obtained. The structure of this *Ramsey* pattern is illustrated in Fig. 5.11. Since the slope of the change in frequency is zero at the peak, the direct measurement of the

resonance frequency from the Ramsey pattern is not suitable for a precise determination. Consequently, the microwave frequency is phase- or frequency-modulated so that an error signal can be generated by synchronous demodulation of the detector response. The frequency stability of the cesium beam standard then depends upon factors such as the modulation used in the microwave interrogation cavity and the frequency locking scheme. A generic block diagram of a cesium standard is illustrated in Fig. 5.12.

The frequency stability is approximately given by

$$\sigma_y(\tau) = \frac{K_{Cs}}{Q_1 \cdot \left(\frac{S}{N}\right) \cdot \tau^{1/2}}, \quad (5.20)$$

where $Q_1 = \nu/\Delta\nu$ is the line quality factor, S/N is the signal-to-noise ratio of the detected signal (the noise is mostly shot noise at the detector) and K_{Cs} is a factor dependent upon the modulation used but close to unity. In a typical, well-built laboratory primary standard a stability of $5 \cdot 10^{-12} \text{ s}^{1/2}/\tau^{1/2}$ over a range extending to 40 days has been measured.

Commercial cesium beam devices are widely available and in use today. However, the wide availability of GNSS timing receivers used to distribute precise time, their superior performance especially coupled with a rubidium clock and most notably their low cost has impacted the cesium frequency standard market. At this time, technology for primary laboratory and precision timekeeping devices has moved into cold atom technology applied to so-called fountain clocks, which will be discussed in Sect. 5.2.3.

Hydrogen Maser Frequency Standards

Hydrogen masers are the most stable frequency standards commercially available for use in laboratory and

ground station environments. They have been developed for scientific, timekeeping and GNSS applications. There are two basic designs of hydrogen masers in use, the active maser where the maser cavity actually oscillates and produces a signal actively [5.17, 18], and the passive maser whose cavity is passively interrogated in a similar manner to the rubidium and cesium devices just discussed [5.19]. A third design of hydrogen maser known as the Q-enhanced maser [5.20] that can operate in either mode is briefly presented in Sect. 5.3.3.

The hydrogen maser operates at the ground state between the two hyperfine levels of atomic hydrogen ($F = 1, m_F = 0$ to $F = 0, m_F = 0$) at the ground state hyperfine frequency, ν_{hf} , of 1420.405752 MHz. The hyperfine energy levels of atomic hydrogen are shown in Fig. 5.13. The transition frequency depends on the magnetic field B (expressed in Teslas) and amounts to

$$\nu = \nu_{\text{hf}} + 1399.08 \cdot 10^7 \text{ Hz} \cdot B^2. \quad (5.21)$$

At room temperature the population of hydrogen atoms is nearly evenly distributed between four magnetic hyperfine levels designated by $F = 1, m_F = 1, 0, -1$ and $F = 0, m_F = 0$. These energy levels depend on the relative orientation of the magnetic dipoles associated with the proton and the electron when the atom is in a magnetic field. In the upper level, designated as $F = 1$, the angular momenta of the proton and electron are aligned and added. Their magnetic dipoles are also aligned. In this state the total angular momentum can orient itself with a magnetic field in three different directions and the $F = 1$ energy level splits into three components. The $F = 0$ energy level results from the alignment of protons and electrons that cancel their total angular momentum and their magnetic dipoles oppose each other.

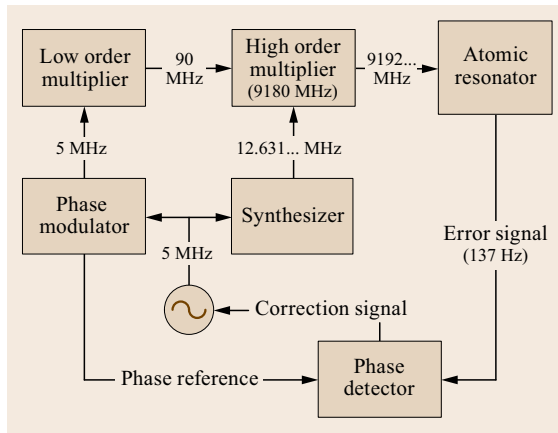


Fig. 5.12 Generic cesium beam standard block diagram

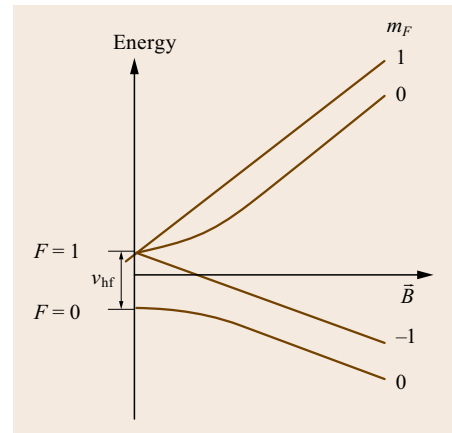


Fig. 5.13 Energy levels of atomic hydrogen

Figure 5.14 shows a schematic diagram of an active hydrogen maser. Molecular hydrogen at a pressure of about 0.1 Torr is dissociated into atomic hydrogen by an RF plasma discharge and collimated into a beam. Atoms in two of the upper magnetic hyperfine energy levels ($F = 1$, $m_F = 1$, and 0) are selected by passing through a highly inhomogeneous magnetic field generated by a multipole permanent magnet, which causes them to move toward the weak field near the axis of the magnet. These atoms are focused into a storage bulb located in a resonant microwave cavity tuned to the atomic hydrogen hyperfine frequency. The storage volume confines the atoms to a region where the oscillating magnetic field is in the same phase.

As the atoms proceed from the multipole magnet into the cavity bulb, the magnetic field they encounter changes from about 9 kGauss radially in the magnet to about one Gauss along the axis of the beam. In this drift region the atoms remain in the $F = 1$, $m_F = 1$, and 0 state and will be kept in these states along the drift region, if the magnetic field they encounter is reduced to the level of the field in the resonator without sudden interruption or change in direction.

A very important feature of an active hydrogen maser is the monomolecular Teflon surface coating in the storage bulb that enables its operation as an oscillator with a narrow resonance line width. This is achieved

by storing the atoms without appreciable loss of phase coherence from collisions with the wall surfaces or each other.

The frequency of the $F = 1$, $m_F = 0$ to $F = 0$, $m_F = 0$ transition that powers the oscillator depends on the magnetic field. To avoid frequency shifts from changes in the magnetic field, hydrogen masers are operated at low magnetic fields. To maintain these low fields and provide a spatially uniform field, with variation at the micro-Gauss level throughout the bulb, magnetic shields are placed about the resonator to attenuate the outside magnetic field, and a solenoid is placed within the innermost shield to provide a uniform and controllable field.

Maser oscillation is sustained when the energy released by the incoming atoms resulting from stimulated emission of the microwave fields in the resonator exceeds the energy lost by the resonator. The energy loss includes the signal delivered to the receiver though a pickup loop in the microwave cavity that is mixed and compared to a signal from the local oscillator to produce the final output signal.

The fundamental stability limit for the maser is similar to other oscillators and is given as

$$\sigma_y(\tau) = \frac{1}{Q_1} \sqrt{\frac{kT}{2P\tau}}, \quad (5.22)$$

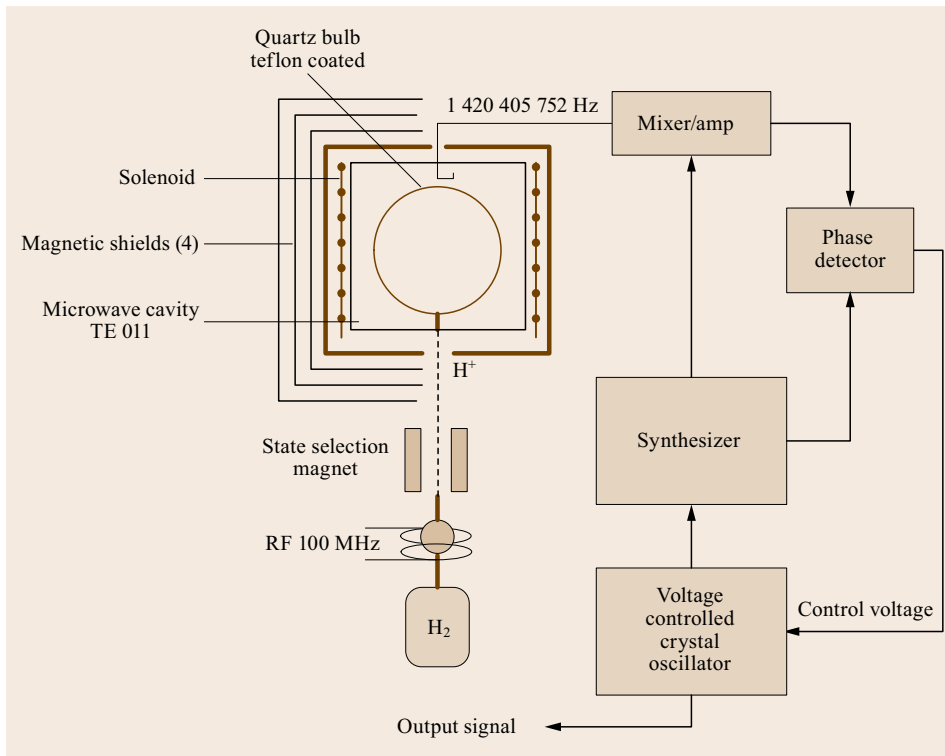


Fig. 5.14 Active hydrogen maser schematic

where Q_1 is the line quality factor of the maser operating at a power level P , k is Boltzmann's constant, and T is the absolute temperature. This expression then implies high values for Q_1 and power. However, high power also promotes increased interatomic collisions. Consequently masers typically operate with low power in which the signal-to-noise ratio of the receiver of the maser signal has a significant effect on the short-term stability (at $\tau < 100$ s).

A typical active maser uses a microwave cavity resonator operating in the TE_{011} mode. Without appreciable dielectric loading by the bulb, the resonator's dimensions are of a cylinder of 28 cm in diameter and height. These dimensions result in a storage bulb of two to three liters in size and a considerable large size to the overall maser with the levels of magnetic shielding and vacuum enclosure required. A reduced resonator size has been achieved by using dielectrically loaded cavities. In active mode, these smaller resonators tend to suffer larger thermal variations in resonance frequency due to the properties of the dielectric material used for loading the cavity and therefore require more thermal control than the unloaded resonators. However, used in the passive mode a considerable size reduction can be achieved. In those cases magnetron cavity designs have achieved a small resonator size with sufficient line Q_1 for maser operation.

The smaller resonators using lumped capacitance loading to reduce dimensions are used in passive hydrogen maser and Q-enhanced hydrogen maser designs. Among others this type of resonator is used in the Russian Ch-176 hydrogen maser, some masers built by the Sigma-Tau Standards Corporation and in the small Q-enhanced spaceborne masers developed at the Hughes corporation for the US Naval Research Laboratory.

The use of a cavity loaded with dielectric material enables a smaller size of the cavity resonator, and is thus an important means for reducing the overall size of a hydrogen maser. However, the penalty is that the cavity Q_1 will be lower and may be beneath the self-oscillating limit. Therefore the maser can be operated in the passive mode with two coupling loops whereby one injects a microwave signal into the cavity at the hyperfine frequency and another is used to detect the amplified signal. The energy from the injected signal causes stimulated emission of the atoms in the cavity. These smaller designs usually operate in the passive mode rather than the active mode just described.

A generic passive hydrogen maser design using a probe signal with two modulated frequencies is illustrated in Fig. 5.15. Here, the maser cavity is interrogated with a probe signal that is phase modulated at two different frequencies, f_1 that corresponds to the nom-

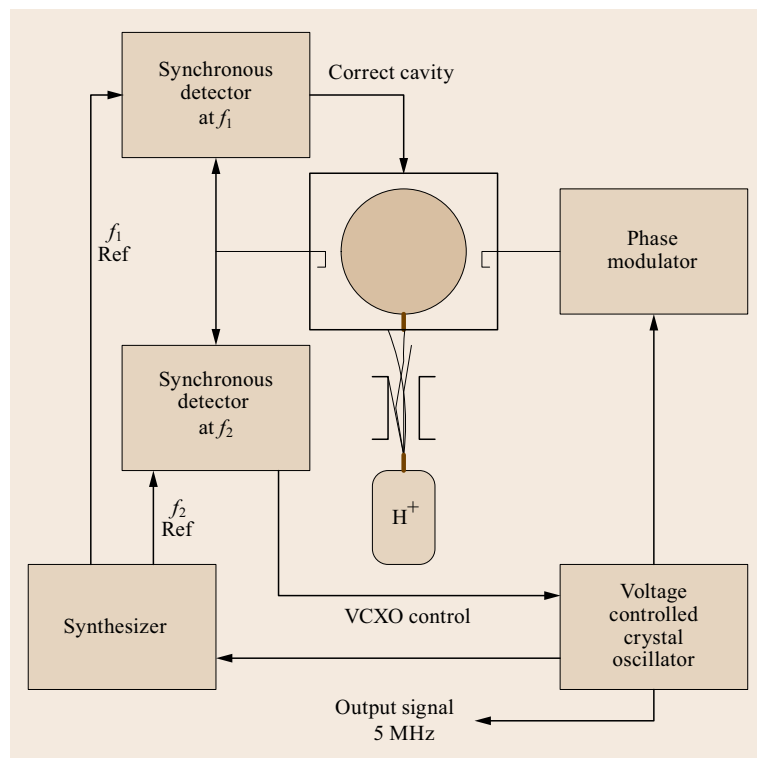


Fig. 5.15 Passive hydrogen maser schematic

inal half-width of the microwave cavity, and f_2 that corresponds to the nominal half-width of the hydrogen resonance. This phase-modulated probe signal is then coupled to the microwave cavity containing atomic hydrogen appropriately state selected. In the resulting spectrum the f_2 sidebands primarily interact with the narrow hydrogen line while the f_1 sidebands primarily interact with the broad cavity resonance. The signal transmitted from the cavity is amplitude modulated at frequencies f_1 and f_2 . The size and sign of the amplitude modulation at f_2 relative to the impressed phase modulation is proportional to the frequency offset between the probe oscillator and the center of the hydrogen line. The microwave signal is envelope detected to recover the f_2 amplitude modulation, which is then processed in a synchronous or phase-sensitive detector reference to the f_2 phase modulation. The resulting error signal is used to correct the probe oscillator so that it is precisely centered on the hydrogen line.

Similarly, the f_1 phase modulation simultaneously probes the cavity resonance causing amplitude modulation of the transmitted microwave signal at f_1 , which is proportional to the detuning of the microwave cavity from the center of the probe frequency. The error signal produced from the f_1 amplitude modulation in a synchronous detector is used to tune the microwave cavity to the probe frequency. This cavity servo technique then effectively stabilizes the mechanical dimensions of the cavity to reduce the effects of the environment, primarily temperature. Consequently, the maser cavity is environmentally stabilized and the interrogation frequency produced by the local voltage-controlled crystal oscillator (VCXO) is locked to the hydrogen resonance. The stability for this type of maser in the short term is not as good as the active maser due to the control servos however in the long term they approach similar performance.

5.2.3 Timescale Atomic Standards

Commercial cesium clocks are still the most prevalent standard for timekeeping in other than national timekeeping centers. Second is the active hydrogen maser that is in limited commercial availability. Both of these commercial devices are expensive with the hydrogen masers being about an order of magnitude more expensive than the cesium. The capability of GNSS timing receivers to disseminate time is increasing and many systems are using them to replace precise frequency and time standards as reference standards in timing applications (Chap. 41).

Within national timing centers, such as the National Institute of Standards and Technology (NIST) in the

United States [5.21], the laser cooled cesium fountain has largely replaced the large thermal beam standards that were used as primary standards for determination of the SI second and contribution to the international atomic time scale. Unlike these other standards, cesium fountain clocks are not commercially available so that each center has built their own version. A number of different cesium fountain clocks are now in use throughout the world [5.22, 23] and in 2012 some 21 timing centers used cesium fountain clocks as their primary frequency standard. These primary standards serve as the metrologic reference and their performance is therefore determined by comparison and coordination with the Bureau International des Poids et Mesures (BIPM) [5.24].

The atomic fountain was first proposed and attempted by Jerrold Zacharias [5.23]. The original objective was to increase the interrogation time of a particular transition in the atoms beyond that possible in a thermal beam device by the use of gravity. If the atoms are launched upward through the same interaction region twice, the Ramsey fringes are produced with a resolution determined by the time between the two interactions. This significantly reduces the sources of errors since the same cavity would be used for both interactions. The design of fountain clocks has become practical through the use of laser cooling of trapped neutral alkali metals with atomic transitions in the microwave range, such as cesium and rubidium.

The development of the magneto-optical trap (MOT) provided the ideal method for trapping a number of neutral atoms and cooling them with laser radiation to within a few hundred micro-Kelvin of absolute zero [5.25]. The balls of cold atoms collected in the MOT could then be launched by the trapping lasers without significant heating and light shifts. This process is illustrated in Fig. 5.16. The atoms in a MOT are confined through the combination of laser field and magnetic field gradients, which can collect large samples of cold neutral atoms from background vapors or atomic beams. The trap geometry contains three intersecting orthogonal pairs of counter-propagating laser beams tuned just below a strong cycling transition in alkaline-earth and alkaline-earth-like atoms. The trapping region produces a three-dimensional optical molasses, so-called because the trap always provides a net force opposing the atom's propagation direction. The addition of a quadrupole magnetic field supplied by a pair of anti-Helmholtz coils forces the atoms toward the center of the trap. Millions of atoms can then be collected in a fraction of a second with an atomic temperature of approximately 1 mK.

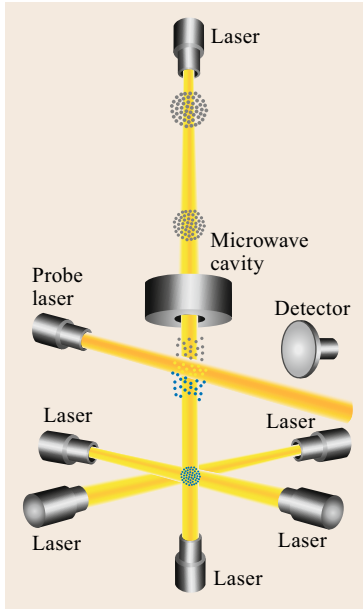


Fig. 5.16 Cesium fountain conceptual diagram (after [5.26]). Image courtesy of NIST

Atomic fountain clocks are based on this concept of laser cooling a collection of atoms to near absolute zero so that these atoms in a nearly-unperturbed neutral atomic state may be interrogated at the ground state hyperfine frequency [5.27]. Although the balls of atoms launched from the MOT contributing to the signal may have a comparatively small number of atoms, on the order of 10^3 to 10^6 , the line quality factor Q_l is so large that a gain in frequency stability is obtained over the conventional thermal beam approach. It is shown that the frequency stability is given by the Allan deviation as

$$\sigma_y(\tau) = \frac{1}{\pi Q_l} \frac{\sigma(\Delta N)}{N} \sqrt{\frac{T}{\tau}}, \quad (5.23)$$

where $\sigma(\Delta N)$ is a variance measure of the fluctuations of the number of atoms from ball to ball, N is the average number of atoms in the balls and T is the cycle duration. In practice it is found that $\sigma_y(\tau)$ is about $3 \cdot 10^{-13} \text{ s}^{1/2} / \tau^{1/2}$, which is in agreement with the expression above.

Development of this technology has been facilitated by the availability of the appropriate lasers for the MOT and interrogation lasers. Diode lasers are the lasers of choice for use in fountain clocks and it is expected that there will be a significant synergy between cold atom development and space technology. The launch and operation of the Atomic Clock Ensemble in Space (ACES) experiment, which incorporates a cold atom cesium clock with a passive hydrogen maser, will demonstrate

the potential of this type of laser-cooled standard in space [5.28].

5.2.4 Small Atomic Clock Technology

The capability of manufacturing very small or miniature atomic clocks was greatly enhanced by the development of a technique known as Coherent Population Trapping (CPT) [5.29]. It makes use of lasers for the optical pumping of clock transitions and been successfully employed with both Rb and Cs. Similar in concept to the classical Rb standard gas cell design, the spectral lamp is replaced by a diode laser at the required D_1 and D_2 wavelengths for Rb of 780 nm and 794 nm, or for Cs at 852 nm and 894 nm. Using diode lasers, the spectrum is much narrower than that produced by the spectral lamp, which improves the pumping efficiency. If the laser signal is modulated to produce two coherent signals at the wavelengths of the optical transitions corresponding to the levels $F = 2$ and $F = 1$ in the case of Rb a new phenomenon takes place. Coherent population optical pumping at the exact resonances with the optical transitions creates an interference resulting in the absence of the absorption of radiation.

The transition energy levels are illustrated in Fig. 5.17. The atoms are trapped in the ground state and find themselves in a nonabsorbing coherent superposition of the two hyperfine ground states. The atomic medium then becomes transparent at the exact resonance of the optical transition. The transmission of the cell increases at resonance and if the frequency of the laser signals is slowly swept, a resonance signal is observed at the photodetector. The shape of the signal is similar to the signals observed in the classical passive Rb standard described above.

In practice, the two laser signals may be obtained from a single laser modulated in frequency at a submultiple of the hyperfine frequency. The highly correlated sidebands created are used to provide the resonance signals. The technique may be used to implement a passive standard in either using the transmission (bright line) or the fluorescence (dark line).

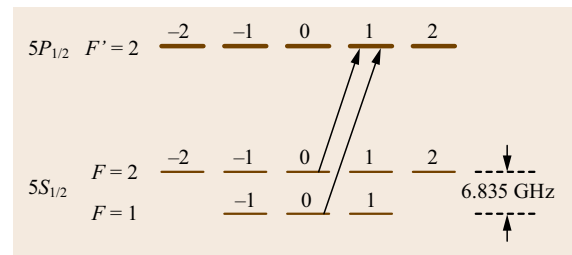


Fig. 5.17 The Rubidium CPT transition energy levels diagram

A microwave cavity is not needed since no microwave signals are required to excite transitions within the ground state. The design of a passive device is shown in Fig. 5.18.

The frequency stability of the CPT passive standard is expressed as approximately the same as that of the intensity optical pumped standard [5.29]

$$\sigma_y(\tau) = \frac{K}{4 \nu_{\text{hf}} q} \sqrt{\frac{e}{I_{\text{bg}} \tau}}. \quad (5.24)$$

Here K is a constant depending upon the type of modulation used and is of the order of 0.2, e is the charge of the electron, I_{bg} is the background current created by the residual transmitted light reaching the photodetector, τ is the averaging time and q is a quality figure defined as the ratio of the contrast C to the line width. The contrast is defined as the CPT signal intensity divided by the background intensity.

The lack of requiring a microwave cavity facilitates the design of miniaturized atomic clocks down to a size limit determined by the laser and its corresponding performance limitations. The application of the CPT technique to cesium has resulted in a commercial product known as the chip scale atomic clock (CSAC) [5.30]. Preproduction units of the CSAC have demonstrated a stability of better than $3 \cdot 10^{-10} \text{ s}^{1/2} / \tau^{1/2}$ at timescales of at least 1–100 s [5.30]. When used in

GNSS receivers, the improved stability helps to reduce phase noise, allows clock coasting during periods of reduced satellite visibility, and supports a faster time-to-first-fix (Chap. 13). Use of a stable atomic clock such as the CSAC inside a GNSS receiver has also been demonstrated as a technique for mitigating wideband radio frequency interference generated by GNSS jamming devices [5.31].

5.2.5 Developing Clock Technologies

Microwave standards are a mature technology and still have good potential for further significant improvements. For instance, a juggling rubidium fountain clock that launches multiple balls of atoms in rapid succession could greatly improve the SNR and result in short-term fractional frequency stability in the high 10^{-15} s at one second while still maintaining excellent long-term systematics well below 10^{-16} . This stability requires a local oscillator with better performance than an OCXO. The Time and Frequency Group at the Jet Propulsion Lab (JPL) in Pasadena has built a cryogenically cooled sapphire-loaded ruby oscillator that achieves $3 \cdot 10^{-15}$ performance from 1–1000 s, thus meeting the local oscillator requirements for an advanced fountain [5.32]. It is possible that further refinements to the fountain concept could bring that device into the low 10^{-15} s at a second.

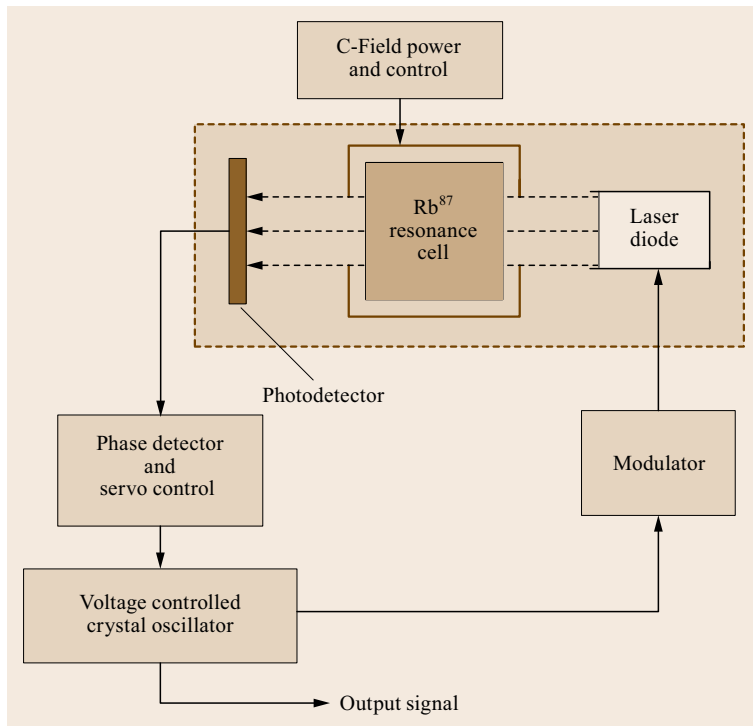


Fig. 5.18 Passive CPT rubidium clock

Laser-cooled microwave ion standards are expected to have an exceptional long-term systematic noise floor. It is likely that the main limitation will be magnetic field sensitivity, which is largely an engineering problem of providing good shielding while still maintaining good optical access. However, while the systematic floor is likely to be in the low 10^{-17} s, the short-term stability is probably limited to the low 10^{-13} s due to the low signal-to-noise ratio inherent in a device with only a few ions. As a result, a microwave laser-cooled ion trap device is unlikely to meet the stated goals.

Buffer-gas-cooled ion standards have already demonstrated a stability of $3 \cdot 10^{-14}$ at one second [5.33]. These devices have large signals (many ions) but only a moderate SNR due to large background signals. A factor of 3–10 improvement in SNR could be achieved with better detection schemes to reduce background. This almost certainly means using lasers instead of lamps as is the current practice. One of these ion standards coupled with an advanced local oscillator (such as the cryogenically-cooled oscillator already discussed) could get close to the short-term stability goal, but the systematic floor is unlikely to be below 10^{-16} (larger numbers of ions at higher temperatures means both exposure to higher RF fields and larger Doppler shifts). Nevertheless, this type of approach should not be dismissed too quickly, since this frequency stability still allows several picoseconds (ps) timing stability at one day. The buffer-gas-cooled ion standard with laser interrogation would require the fewest technological advances and would be the simplest to implement.

The next step in atomic clock evolution is to move from microwave clock frequencies to optical frequencies [5.34–36]. With frequencies measured in the 10^{15} Hz range instead of 10^{10} Hz, optical clocks have a potentially huge gain in line quality factor Q . Since

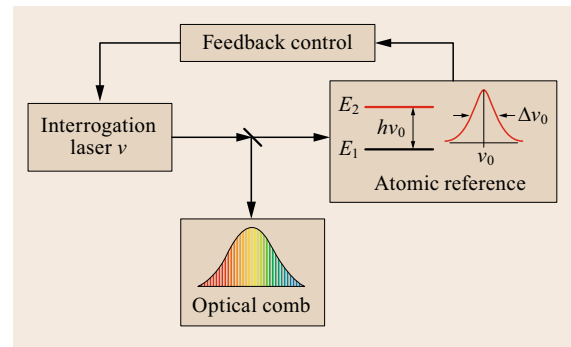


Fig. 5.19 Diagram of optical clock

short-term stability is inversely proportional to Q , it also improves. An ion trap clock based on an optical transition then combines very good short-term stability due to the high Q of the optical transition with an exceptionally low systematic noise floor. An optical clock is shown schematically in Fig. 5.19.

The two technologies that are critical to optical clock progress are the optical comb and laser frequency stabilization. The optical comb (also known as the frequency comb) enables the coherent translation from optical frequencies to microwave frequencies where timing information is usually used, transferred and analyzed. This is a huge step for optical clocks, since previous chains linking optical to microwave frequencies required man-years of highly skilled work and large amounts of equipment to build and maintain. The second critical technology is laser frequency stabilization. To take full advantage of the optical line quality factor, the clock laser, which is now the local oscillator for this clock, must have a frequency uncertainty on the order of 1 Hz or less. This is difficult to achieve but offers great potential for future development of clock technology.

5.3 Space-Qualified Atomic Standards

The development of space-qualified atomic clocks has its origin in the navigation satellite concepts of the late 1960s and early 1970s. The Transit Doppler navigation system [5.37] first demonstrated the potential of worldwide high-accuracy navigation by satellite. These early navigation satellites were in a low altitude orbits, which provided sufficiently strong signals for users to calculate their position from the observed Doppler shift. The oscillator, or clock, had to be stable enough to permit a good frequency measurement over the period of time the satellites were in view. If the oscillator had a frequency change within that interval the frequency

measurement would be in error creating a position error as well.

The advanced navigation satellite system designs, such as the Naval Research Laboratory (NRL) TIMATION (time navigation) concept and ultimately the Global Positioning System (GPS) were based on passive ranging to provide continuous accurate navigation to their users [5.38]. Development of space-qualified clocks for GPS was focused on predictable stability over time intervals of typically a day. From NRL work in the first phase of the GPS program (Block I) a space clock development program was formed to

develop space-qualified clocks for the NAVSTAR satellites [5.39]. To meet the system error requirements, projects in rubidium, cesium and hydrogen maser units were initiated. Improvements to the rubidium and cesium units used in the Block I demonstration satellite constellation were required to support the producibility, reliability and performance needs of the operational satellites (Block II/IIA) and alternative industrial sources for these units were required to be developed. The Block II/IIA GPS satellites contained two space cesium clocks and two rubidium clocks. Subsequent blocks of satellites, the replacement Block IIR and improved Block IIR-M contained three rubidium clocks and Block IIF contains two space rubidium and one space cesium clock. The next Block III satellites are to contain three rubidium clocks.

Developments of space-qualified atomic clocks have also been conducted for the Russian Federation GLONASS (global navigation satellite system), the European Galileo system, and the Chinese BeiDou navigation satellite system. Each GLONASS and GLONASS-M satellite hosts Russian three Cs-beam frequency standards [5.40], whereas the latest generation of GLONASS-K1 satellites is equipped with both cesium and rubidium clocks [5.41]. The Galileo satellites make use of passive hydrogen masers as their primary clocks in addition to conventional Rb gas cell frequency standards [5.42]. China, finally had a development project into space rubidium clocks since the initial deployment of BeiDou [5.43, 44] and is also investigating use of hydrogen masers for their global navigation systems [5.45].

Overviews of spaceborne atomic frequency standards and their use in the individual global and regional navigation satellite systems are given in [5.46] and [5.47]. The specific design aspects that distinguish spaceborne clocks from their terrestrial counterparts (such as environmental robustness and utmost reliability) are further discussed in [5.48].

5.3.1 Space Rubidium Atomic Clocks

The first atomic clocks in orbiting satellites were flown on board the Navigation Technology Satellite one (NTS-1) [5.49]. This satellite contained two quartz crystal clocks, developed under the TIMATION program, and two experimental rubidium clocks based on the FRK unit built by Efratom of Munich [5.50]. These rubidium clocks were commercial clocks modified as an experiment to survive the launch and thermal environment of space. Both rubidium units were encased in a large radiation shield to reduce the effects of radiation on the clock electronics. The performance of the NTS-1 rubidium clocks is shown in Fig. 5.20 in com-

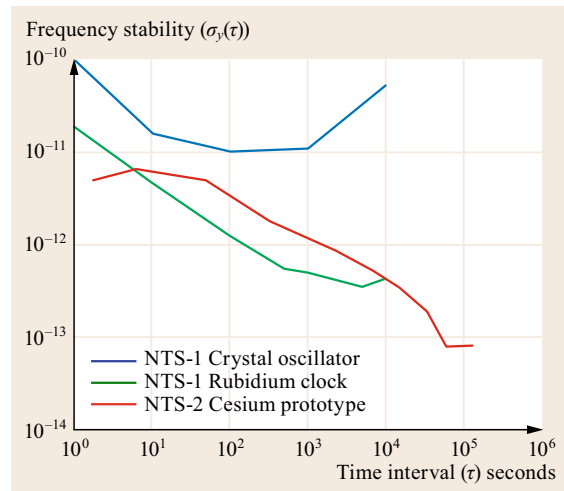


Fig. 5.20 Performance of NTS space clocks

parisons with the NTS-1 quartz oscillators and a cesium frequency standard flown later on the NTS-2 spacecraft. Despite a notable frequency drift at time exceeding several hours, these units provided the proof-of-concept necessary for use of rubidium clocks as primary units in the first NAVSTAR developmental satellites.

The rubidium atomic clocks used in the early Block I GPS satellites were built by Rockwell International based on the FRK design of Efratom [5.51]. These units were a combination of an electronics design rebuilt for use in space by the Anaheim Division of Rockwell International and the physics portion built by Efratom. Fig. 5.21 shows the qualification unit of this design.

The early performance of the GPS units had a number of difficulties but they provided sufficient performance to support continued development. Nevertheless, the atomic clock of choice – and what was perceived to be the best clock for the final operational system (GPS Block II) – became the space-qualified cesium beam clock (Sect. 5.3.2), since it was a primary standard and did not exhibit the significant drift characteristic typical of rubidium. During the development of the GPS operational system in the 1980s alternative space-qualified atomic clocks were developed as part of the operational program of deployment [5.39]. The alternative source for space-qualified rubidium clocks was a design originally from EG&G, who became Perkin Elmer Optoelectronics and are currently known as Excelitas [5.52]. That company produced two prototype units that went on to become the atomic clock of choice for the GPS system during the Block IIR satellite deployment. Developments for GPS satellites focused on performance, improvements in the clock's state-of-health diagnostics, ground testability, and reduction of environmental sensitivities.



Fig. 5.21a–c Space-qualified rubidium clocks for GNSS Satellites: a qualification model for the early GPS satellites (a) (image courtesy of NRL), the RAFS developed by TEMEX/Spectratime for Galileo (b) (image courtesy of Spectratime), and the inside view of a second generation BeiDou rubidium clock (c) (after [5.44])

The latest version of Excelitas rubidium frequency standards used on board the GPS Block IIF satellites [5.53] as well as the *Michibiki* satellite of the Japanese quasi-zenith satellite system (QZSS) offer roughly a factor-of-two improvement in noise level and offer a stability of about $\sigma_y(\tau) = 1 \cdot 10^{-12} \text{ s}^{1/2}/\tau^{1/2}$. This is mainly achieved through the use of a xenon buffer gas and an advanced filter for the rubidium spectral lines that increase the overall quality factor [5.54]. Both the Block IIR and Block IIF operational rubidium clocks have demonstrated outstanding inflight performance as discussed further in Sect. 5.3.6.

Even though several BERYL Rubidium clocks [5.55] were flown on GLONASS precursor satellites [5.46], the Russian navigation has focused on the exclusive use of cesium beam frequency standards for more than 30 years (Sect. 5.3.2). Only recently, rubidium clocks have been introduced as alternative atomic frequency standards in the GLONASS-K series [5.41]. However, no flight results have become available up to the end of 2015.

Rubidium atomic frequency standards (RAFS) for the European Galileo program were originally developed under Swiss management by Spectratime (formerly Temex Neuchatel Time) along with Astrium, Germany, who did the space-qualified electronics [5.56]. The clocks exhibit representative stabilities of $\sigma_y(\tau) = 2\text{--}4 \cdot 10^{-12} \text{ s}^{1/2}/\tau^{1/2}$ and were flight tested in the GIOVE-A and -B satellites [5.57] prior to their selection and incorporation into the operational Galileo satellites. A sample of the Galileo RAFS is shown in Fig. 5.21. Complementary to the Galileo program, a variant using a Swiss electronic package has also been developed for use as backup clocks within the Chinese BeiDou constellation [5.47, 56]. Furthermore, the Spectratime Rubidium clocks are employed as primary frequency standards for the Indian Regional Navigation Satellite System (IRNSS;

now known as NavIC for Navigation with Indian Constellation).

Along with the buildup of their national navigation satellite systems, various types of space-qualified rubidium clocks have also been developed in China. These indigenous RAFS exhibit a reported stability of about $5 \cdot 10^{-12} \text{ s}^{1/2}/\tau^{1/2}$ and presently serve as primary onboard clocks for the regional BeiDou navigation system. A recent RAFS model developed by the Beijing Institute of Radio Metrology and Measurement is shown in Fig. 5.21.

5.3.2 Space-Qualified Cesium Beam Clocks

The first prototype cesium clocks evaluated in orbit for GNSS were contained in NTS-2, which was the precursor of the NAVSTAR Block I satellites [5.38, 39, 58]. Two prototype cesium units were flown in NTS-2 and provided the space qualification of the cesium tube necessary for continued development. The cesium tube qualified in NTS-2 developed by Frequency and Time Systems Inc. (FTS) is shown in Fig. 5.22 and was the same tube used in the operational Block II/IIA GPS satellites. Performance of the NTS-2 cesium units in orbit is also shown in Fig. 5.20.

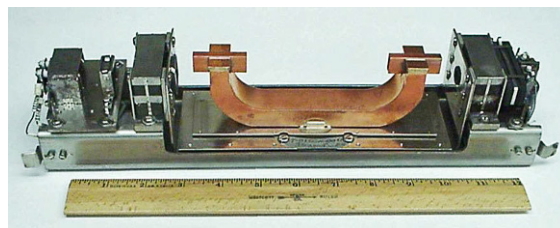


Fig. 5.22 Cesium beam tube without external shields and vacuum container showing the Ramsey cavity, the cesium reservoir on the right and detector assembly on the left. Image courtesy of NRL

During the next development step, engineering models of a refined design were built and tested. In cooperation with the US Defense Nuclear Agency, complete radiation testing was performed to determine the design parameters necessary for a radiation-hardened unit. The unit design and development with FTS continued through the preproduction model (PPM) stage. Six of these PPMs were built and provided to the prime satellite contractor Rockwell International (RI) for use in the early NAVSTAR satellites. The first PPM was launched in NAVSTAR 4 and the last would have been launched in NAVSTAR 7. This cesium design became the one employed in the Block II and IIA operational GPS satellites. These satellites had two cesium and two rubidium clocks in each satellite.

The next generation of cesium clocks for space were developed and deployed in the GPS Block IIF satellites [5.53]. These clocks employ a similar thermal cesium beam tube to that used in the earlier GPS satellites, but use digital electronics rather than the analog electronics used previously. The complete unit of the digital cesium beam frequency standard (DCFBS) for GPS Block IIF is shown in Fig. 5.23. It achieves a representative stability of $1 \cdot 10^{-12} \text{ s}^{1/2}/\tau^{1/2}$, and is mainly used as backup clock on those satellites. Its inflight performance is further described in Sect. 5.3.6.

Aside from GPS, cesium beam atomic clocks are also used extensively within the Russian GLONASS constellation, where they have served as the primary source of time and frequency information on most satellites launched so far. Satellites of the first generation GLONASS satellites were equipped with three GEM clocks [5.55] built by the Russian Institute of Radio Navigation and Time (RIRT), formerly Leningrad Scientific Research Radiotechnical Institute

(LSRRI), in St. Petersburg. Only one clock is active at a time, while the others are kept in cold redundancy [5.40]. Even though the GLONASS clocks were operated in a sealed compartment at ambient pressure and temperature, their limited survivability posed a major constraint to the overall lifetime of those satellites [5.60].

Satellites of the subsequent GLONASS-M series, which still makes up the majority of the current constellation, are equipped with MALAKHIT clocks that are likewise built by RIRT. Following [5.55, 61], the two clock types exhibit Allan deviations of about $1.5 \cdot 10^{-10} \text{ s}^{1/2}/\tau^{1/2}$ and $3 \cdot 10^{-11} \text{ s}^{1/2}/\tau^{1/2}$ at timescales of 100 s to one day. More recent developments have resulted in a performance improvement by a factor of 2–3 as well as a notably reduced mass of the GLONASS on-board frequency standards [5.62].

5.3.3 Space-Qualified Hydrogen Maser Clocks

The use of hydrogen masers for the ground stations and eventually in spacecraft was considered for GPS before the beginning of the program. Efforts at that time were based on adapting the active hydrogen maser design developed by the Smithsonian Astrophysical Observatory (SAO). SAO developed a series of active hydrogen masers for the Very Long Baseline Interferometry program that were capable of being operated in ground stations at remote sites. From that design, SAO built a space-qualified hydrogen maser for the National Aeronautics and Space Administration's (NASA) Gravity Probe One unit to investigate the gravitational relativistic effects on a precise clock. The probe was launched in the mid-1970s in a vertical ballistic trajectory to maximize the relativistic effects on the atomic clock [5.63]. The successful flight qualification and operation in the 2.5 h launch profile demonstrated the potential for operating such a clock in orbiting spacecraft. However, this particular active hydrogen maser design was rather large to consider incorporating into an orbiting satellite.

Reduction in the size of such a space clock was considered essential, so compact passive physics unit designs were investigated by NRL for GPS [5.64]. Various passive maser designs were investigated, both of the physics unit as well as electronics designs for cavity stabilization and interrogation. Several experimental units were built incorporating design alternatives for evaluation and the most successful approach was the Hughes Q-enhanced design incorporating a small magnetron cavity [5.65, 66]. This compact passive hydrogen maser design reduced the overall size of the device to roughly the size of a GPS space-qualified cesium clock.

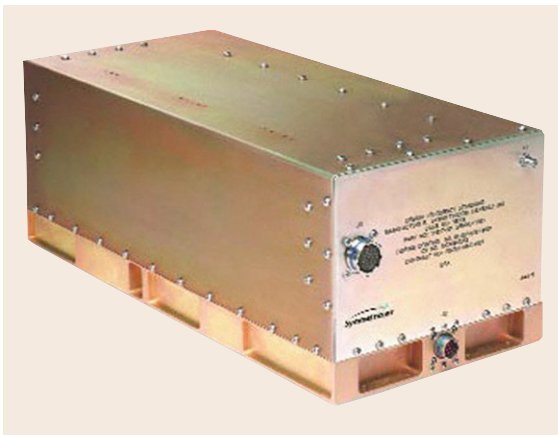


Fig. 5.23 GPS Block IIF digital cesium beam frequency standard (DCFBS) from Symmetricom (after [5.59]). Reproduced with permission of MicroSemi

Figure 5.24 shows the physics unit of the final version of the Hughes design.

The Galileo program also developed a small passive hydrogen maser (PHM) clock for operational satellites [5.67–69]. The development was jointly performed by Spectratime, Switzerland, and Galileo Avionica, Italy, who were in charge of the physics package and the electronics package, respectively. The maser cavities used in these devices are of similar design to that of the Q-enhanced maser design discussed above. They are a magnetron cavity design using a metal cavity machined to hold the hydrogen containment bulb with three arms that provide the capacitive loading on the cavity. The microwave cavity is thermally controlled to exhibit variations at the level of a few milli-Kelvin for baseplate temperature changes of about ± 5 K. The Galileo PHM achieves a typical performance of $1 \cdot 10^{-12} \text{ s}^{1/2} / \tau^{1/2}$ over timescales of 1–10 000 s, which marks a notable performance increase over the Galileo RAFSs and makes it one of the best clocks ever used in navigation satellites. On the other hand, the mass of about 18 kg is notably larger than that of the rubidium clock. A flight-qualified space passive hydrogen maser ready for thermal vacuum testing is shown in Fig. 5.25. The PHMs for the Galileo program were flight tested on board the GIOVE-B technology demonstration satellite [5.57] and are now in routine use on board the operational Galileo satellites.

In parallel to the Galileo PHM developments, Observatoire de Neuchâtel and Spectratime pursued the development of a space-qualified, active hydrogen maser for use within the Atomic Clock Ensemble in Space (ACES [5.28]). The Space H-Maser (SHM) uses a sapphire loaded microwave cavity and achieves an Allan deviation down to $5 \cdot 10^{-15}$ at a timescale of 100 s [5.70, 71]. The larger mass (35 kg) and power con-

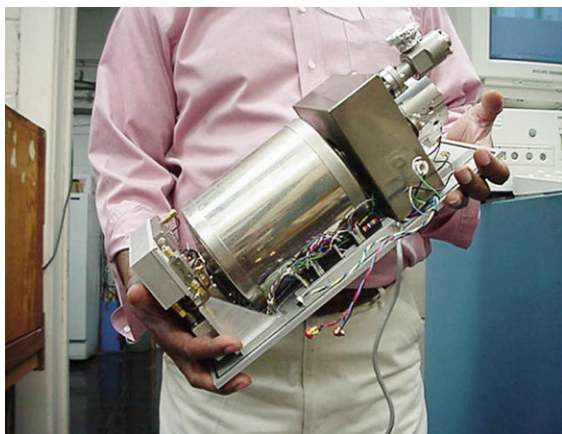


Fig. 5.24 Hughes Q-enhanced hydrogen maser physics unit. Image courtesy of NRL

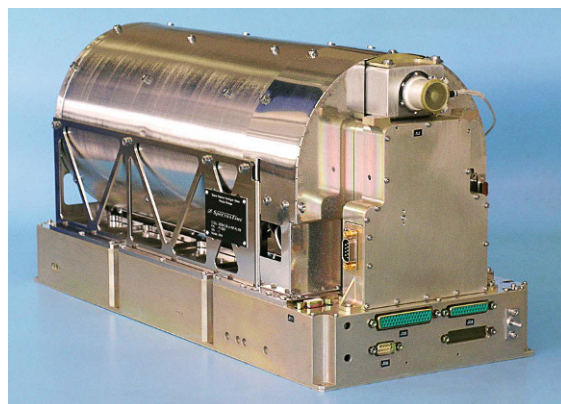


Fig. 5.25 Galileo space passive hydrogen maser. Image courtesy of Spectratime

sumption (77 W) of the active maser did not allow its consideration for the present Galileo system. Nevertheless, its use on ACES will provide further evidence for the potential of high-performance clocks in future navigation systems.

5.3.4 Space Linear Ion Trap System (LITS)

The Jet Propulsion Laboratory Time and Frequency Group have developed a new technology standard known as the linear ion trap standard (LITS) [5.72]. Operational versions of these units are being deployed in the NASA Deep Space Network as replacements for the large active hydrogen masers currently in use. A spacecraft version of these units has been investigated and offers the potential of a very small size and power clock with potentially high stability. The physics package is small but since it is a passive device a high quality local oscillator is needed to gain the full potential of these devices. The potential performance gain using a modest performance local oscillator and the adaptability to digital implementation of the electronics could be a major step in spacecraft clocks [5.73]. NASA is currently developing a space-qualified version of this device for demonstration in a space environment [5.74].

5.3.5 Satellite Onboard Timing Subsystems

Aside from the atomic frequency standard, navigation satellite systems commonly employ some form of frequency distribution unit as part of their timing subsystem. Following [5.75], this unit may serve up to three purposes:

- Selection of one out of multiple clocks as the main source for the time and frequency generation

- Conversion of the native clock frequency to the base frequency for the navigation signal generation, and finally
- Performance of fine frequency adjustments to keep deviations of the onboard time from the GNSS time scale within specified limits.

In advanced timing system implementations, the above functions are combined with a monitor that compares the active clock against a reference to identify potential anomalies such as occasional bad points or outliers, phase jumps and frequency steps. All of these anomalous effects may happen singly, in combination, suddenly, or over a period of time. Serious situations related to satellite clock anomalies can be avoided by detection of these anomalies on board rather than through detection by tracking data on the ground. The clock's behavior can be better monitored on board in real time without additional noise or errors added by the communication link. However, multiple operating frequency standards on board are necessary to accomplish this result.

A well-known example of a timing subsystem taking care of the above functions is the Time Keeping Sys-

tem (TKS) of the GPS Block IIR satellites [5.76–78]. The TKS was originally designed to provide a common interface to different types of atomic clocks as well as determine the differences between the onboard atomic clocks and the output voltage-controlled crystal oscillator (VCXO). The system was configured to provide an interface for three atomic clocks, any one of which, when operating, was compared with a redundant VCXO by a phase comparator running at 600 MHz (Fig. 5.26). The VCXO produces the final signal but is adjusted or disciplined to the atomic clock's output. This inter-comparison produces a measure of the onboard atomic clock's performance but it is ambiguous as to whether it occurs in the atomic clock or in the VCXO. Either one can affect the resultant comparison. At least three clocks would be necessary to be intercompared in order to produce an unambiguous determination of which clock produced the unacceptable performance [5.5].

In a similar fashion to GPS, the Galileo satellites utilize an onboard system called the clock monitoring and comparison unit (CMCU) [5.79–81] to monitor a standby clock with the clock driving the satellite transmitter (Fig. 5.27). Each Galileo satellite is equipped with two rubidium atomic frequency standards as well

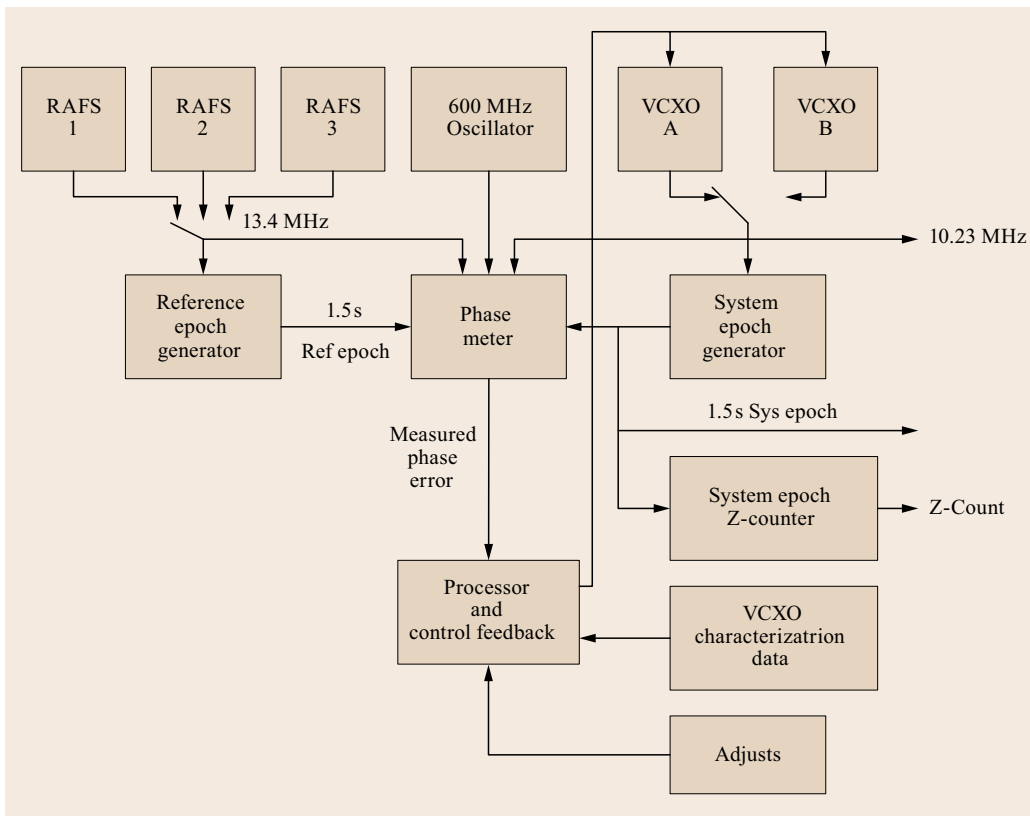


Fig. 5.26 Block diagram of Block IIR satellite Time Keeping System

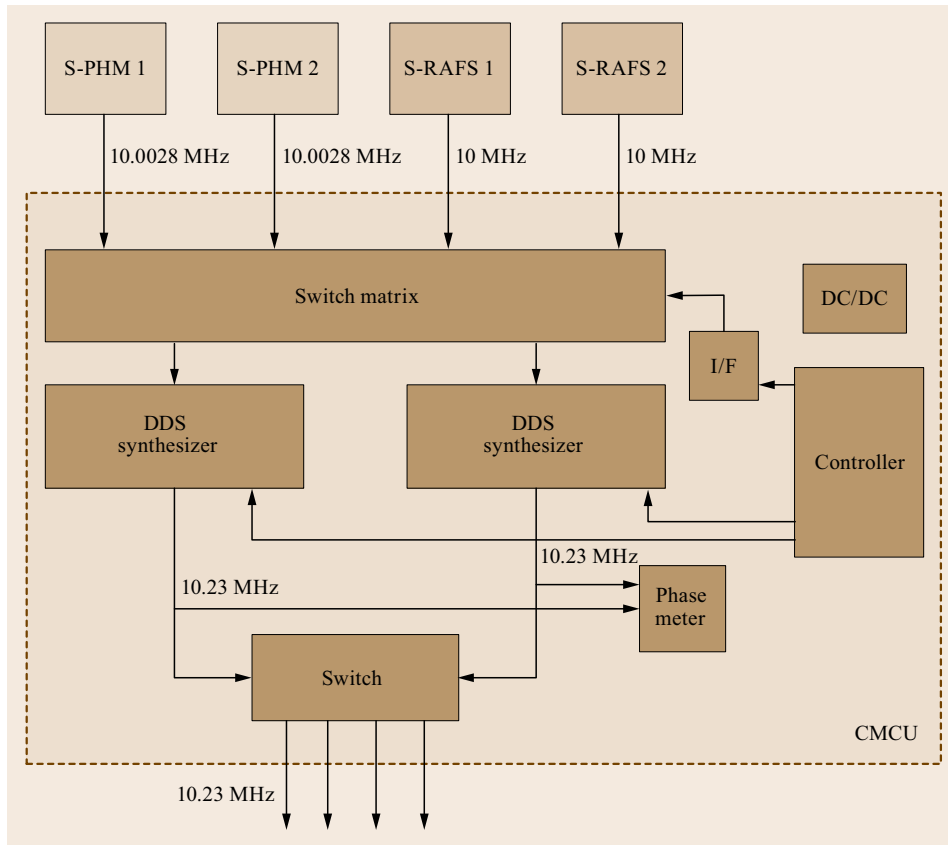


Fig. 5.27 Diagram of the Galileo clock monitoring and comparison unit

as two passive hydrogen masers. At any time, two of these four clocks are active while the other two serve as cold redundancy. Using a switching matrix, the signals of the two active clocks are connected to two synthesizers that shift the native clock frequency of 10 MHz for the RAFSs and 10.0028 MHz for the PHM by roughly 230 kHz to obtain the 10.23 MHz core frequency for the navigation signal. The synthesizers can be digitally controlled and enable adjustments of the output frequency in steps of less than 10^{-15} [5.81]. Even though only one synthesizer is selected to serve as master time reference, both outputs are continuously monitored through a phase meter. By continuously comparing the master clock with an operating standby clock, a switching transient and a loss of knowledge of the satellite time is avoided in case an immediate switch between both clocks needs to be made.

The onboard systems in use today are still mostly dependent upon the performance of a single clock to produce the desired performance. If clocks could be compared on board or with the signals from neighboring satellites provided satellite cross-links or signals from the other satellites could be made available, their outputs could be monitored to be less susceptible to in-

terruption or anomaly but also to produce a somewhat more stable and accurate signal. This onboard comparison capability could provide an immediate detection of anomalies in the operational clock and if properly instrumented possibly even the navigation payload. The resulting status could be inserted into the navigation message for direct broadcast to the users in case of anomaly and to the ground monitoring stations, thereby providing a real-time alerting capability to the system. Data associated with the comparative indication could also be telemetered to the control segment for diagnostic and remedial actions. The current GNSSs do not perform a comparison between the onboard clocks sufficient for determination of clock anomalies. Instead they rely on ground monitoring for detection and correction of such anomalies. A more sophisticated measurement system would be necessary to support a technique for automatic detection and correction on board.

5.3.6 On-Orbit Performance of Space Atomic Clocks

The performance of atomic clocks on board the GNSS satellites is critical to the ultimate accuracy achievable

by the system. Today's satellite navigation systems operate primarily in a passive mode, where the satellites are tracked by a set of monitoring stations. From these observations the system parameters, i. e., satellite ephemerides and clock correction values, are predicted ahead. The navigation information computed at the system master station is then uploaded into the satellites for transmission to the users over time [5.82]. Consequently, the ability to evaluate the clock performance in orbit and to accurately predict the system parameters is required. Effects contributing to the clock observations must be carefully modeled or eliminated, so that the observed and predicted satellite clock information will be as accurate as possible [5.83, 84].

Similar to GNSS-based positioning, the monitoring of GNSS satellite clocks makes use of pseudorange and carrier-phase observations. Both of these reflect measurements of the difference between the signal receive and transmit times relative to the local receiver and transmitter clocks. As indicated by the terminology, pseudoranges (and likewise the carrier-phase observations) do not represent a pure measurement of the distance ρ , but include contributions of the respective clock offsets relative to a common system timescale. As discussed in more detail in Chap. 19, the pseudorange p_i and carrier-phase observations φ_i on the i -th frequency ($i = 1, 2, \dots$) can be modeled as

$$\begin{aligned} p_i &= \rho + c(dt_r - dt^s) + I_i + T + e_i \\ \varphi_i &= \rho + c(dt_r - dt^s) - I_i + T + A + \varepsilon_i, \end{aligned} \quad (5.25)$$

where c is the vacuum speed of light, ρ is the distance between the satellite antenna and the user receiver's antenna, dt_r is the clock synchronization offset of the user receiver's time, dt^s is the clock synchronization offset of the satellite's time at the time of transmission, I_i is the frequency dependent propagation delay due to the ionosphere and T is the delay due to the neutral atmosphere, mostly the troposphere. The measurements errors e_i and ε_i exhibit standard deviation at the decimeter and millimeter level, respectively, for pseudorange and carrier-phase observations. Even though carrier-phase measurements exhibit an extremely low noise, they also include an ambiguity A , which comprises both integer multiples of the wavelength and fractional-cycle phase biases. During uninterrupted tracking of a satellite signal the ambiguity is constant thus offering highly precise measurements of the range change and clock offset variations over time.

The various delays and errors associated with the observations are discussed in depth in Chap. 19. However, once the observations are corrected for instrumentation, equipment effects (antenna offsets and the like), propagation effects, geometric effects or delays, and satellite position relative to the receiving equipment are all compensated, the residual differences

$$\begin{aligned} cdt^s - cdt_r &= (\rho + I_i + T) - p_i + e_i \\ cdt^s - cdt_r - A &= (\rho - I_i + T) - \varphi_i + \varepsilon_i \end{aligned} \quad (5.26)$$

between the measurements and the modeled geometric range and propagation delays are basically a compari-

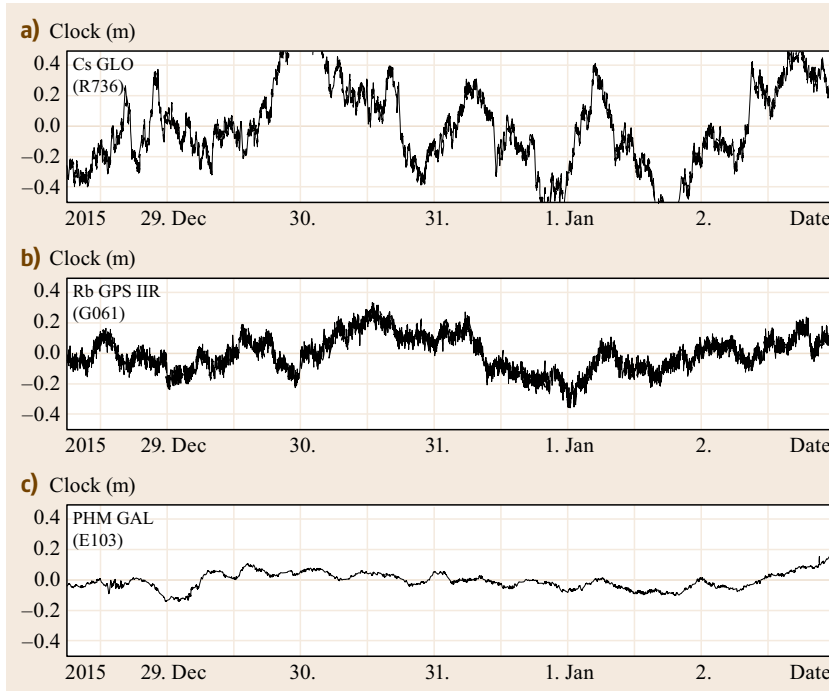


Fig. 5.28a–c Time series of observed clock offsets ($cdt^s - cdt_r$) for selected GNSS onboard frequency standards: GLONASS cesium clock (a), GPS Block IIR rubidium clock (b), and Galileo passive hydrogen maser (c). All values are referred to a highly stable ground clock (active hydrogen maser) and have been detrended with a second-order polynomial. Values in brackets denote the space vehicle number of the respective GNSS satellites. Based on data from a multi-GNSS orbit and clock solution [5.85] of GeoForschungsZentrum (GFZ), Potsdam

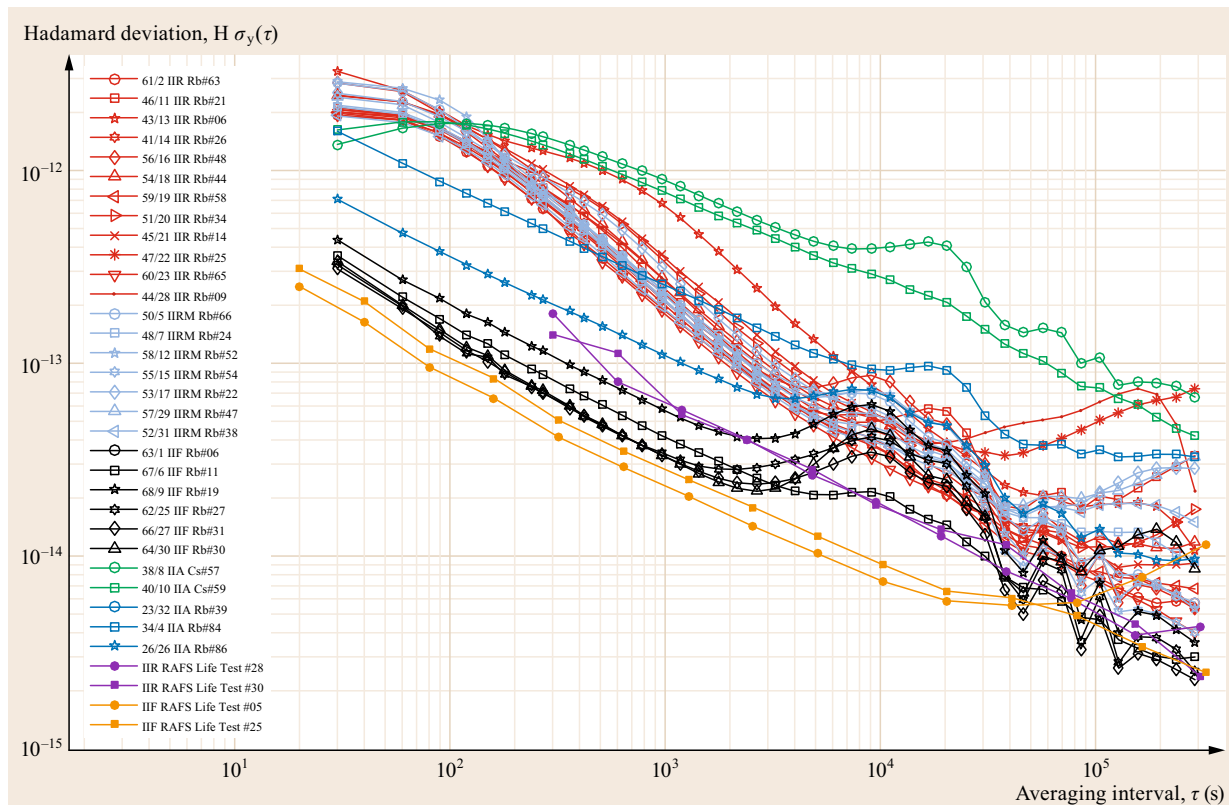


Fig. 5.29 On-orbit performance of GPS satellite clocks (Oct.–Dec. 2014). Individual satellites are identified by their space vehicle number (SVN)/pseudorandom noise (PRN) number. In addition the Block type and the active clock are indicated

son of two clock outputs just as they would be measured in a laboratory evaluation of two clocks. Examples of observed satellite clock offsets relative to a ground reference are shown in Fig. 5.28 for different GNSS satellites. The time series clearly reveal the different short- and long-term stability of individual types of atomic frequency standards.

Characterizing the satellite clocks' performance and predicting their performance ahead is dependent upon the ability to separate all other errors from the measurement so that only clock errors remain. This function is performed for the GNSSs by their respective tracking networks to support their operation. For GPS the network consists of the GPS Ground Segment tracking stations (Chap. 7) combined with additional stations operated by the National Geospatial-Intelligence Agency (NGA). This network supports the real-time operation of the GPS as well as providing the data for producing precise ephemerides for all the GPS satellites. In their support to the science community, the International GNSS Service (IGS; see Chap. 33) performs a similar function to provide highly accurate GNSS orbit products. The IGS data products have significantly

reduced errors over that expected by normal navigation users. Details of the IGS product generation and the processes used to estimate satellite orbits, atmospheric parameters, site coordinates and Earth rotation parameters from a globally distributed network of monitoring stations are discussed in Chap. 34.

Clock offsets estimated from a global monitoring network as part of the precise orbit and clock determination process form a primary means for the inflight performance assessment of atomic frequency standards. In accord with the typical data rates and data arcs used in such adjustment processes, they mostly provide information on the clock stability at timescales of about 5 min to 1 day. For use at short timescales, the method is less suitable though, due to the high computational effort required for high-rate clock solutions. As an alternative, the *one-way carrier-phase* technique (OWCP) has been proposed in [5.86]. It makes use of carrier-phase observations from a single monitoring station connected to a highly stable reference clock (typically an active hydrogen maser). Based on (5.26) the difference of the onboard and ground clock is evaluated using a coarse orbit model such as broadcast ephemerides.

Subsequently, the resulting time series is detrended using a low-order polynomial to remove the impact of the carrier phase ambiguity and residual orbit errors as well as uncompensated atmospheric delays. Subject to a short data arc (up to a few hundred seconds) and the absence of large ionospheric variations, the method can even be applied with single-frequency observations rather than a more noisy dual-frequency combination.

Practical results and a comparison of the OWCP method with other clock estimates have, for example, been reported in [5.87–89]. The unique potential of this method to study the onboard clock stability at subsecond timescales has been demonstrated in [5.90] using GNSS receivers with data rates of 50 Hz and beyond. A special variant of the OWCP method has, furthermore, been studied in [5.91]. It makes use of the triangulation concept [5.92] for a statistical characterization of the clock noise, by processing data from three satellites simultaneously. This method does not require a highly stable ground clock, but relies on equal noise properties of all involved satellite clocks.

The space-qualified atomic clocks in the GPS operational satellites have stability requirements ranging from $2 \cdot 10^{-13}$ /day for cesium and $1 \cdot 10^{-14}$ /day for the rubidium on the later satellites. On-orbit performance has provided better than expected stabilities as demonstrated in Fig. 5.29. This figure provides the frequency stabilities of the GPS constellation clocks as measured

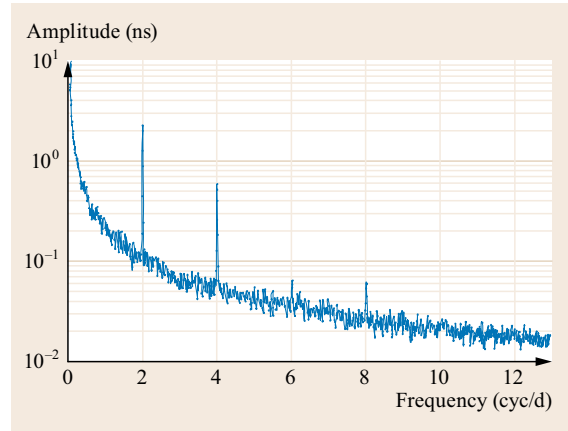


Fig. 5.30 Averaged amplitude spectrum of GPS constellation clocks (after [5.93])

by the Hadamard deviation calculated using the International GNSS Service (IGS) final clock products over the month of November 2013. For comparison with the on-orbit clock performance, Fig. 5.29 also shows the data from two Block IIF rubidium units under long-term test by the NRL. As the plots show, on-orbit performance of the Block IIF units is near that of the environmentally controlled ground units, at least over the short term.

Over longer averaging intervals the performance is degraded in comparison by apparent fixed period

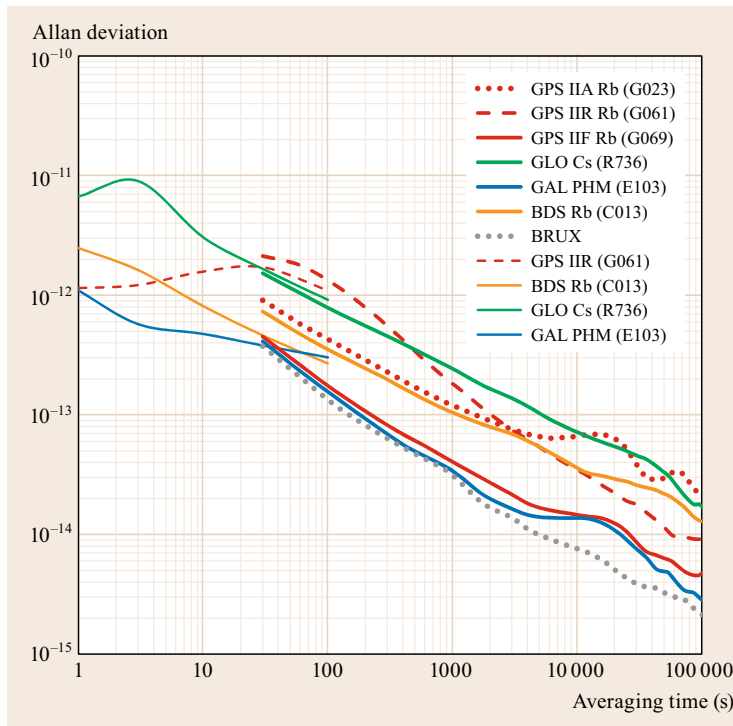


Fig. 5.31 Comparison of atomic clocks performances for current GNSS satellites. *Bold lines* provide Allan deviations for timescales of 30 s to 100 000 s based on data from the clock product of GeoForschungsZentrum (GFZ), Potsdam [5.85] for a one-week data arc in Jan. 2016. A *dotted gray* line provides the corresponding values for a hydrogen maser at the IGS GNSS station in Brussels. For comparison, *thin lines* show clock stability values over short timescales (1–100 s) as derived in [5.90] from a OWCP analysis

harmonic variations attributable to the satellite timing signal [5.93, 94]. The strongest of these harmonic variations occur nominally at 2.003 and 4.006 cycles per solar day with amplitudes of more than 2 ns for some GPS satellites. The Hadamard deviation statistic has a broad frequency response to pure harmonics so the harmonic variations are seen more distinctly using frequency domain techniques, as shown in Fig. 5.30. These data used individual satellite data spectra calculated by applying a standard periodogram with Blackman–Harris windowing to approximately 150 days of IGS final clock data for each satellite referenced to IGS time. The clock data for each satellite were detrended by fitting and removing a second-order polynomial prior to calculating its periodogram. The averaged spectrum was then obtained by averaging the individual satellite spectra at each Fourier frequency. Evidence is increasingly strong that the origin of the variations is likely to be thermal sensitivity, possibly of the associated electronics or of the units themselves.

A comparison of GPS atomic frequency standards with those of other navigation satellite systems in

shown in Fig. 5.31 for timescales ranging from 1 s to more than a day. Stabilities for the various clock types differ by up to a factor of ten and are generally better for rubidium clock and hydrogen masers than for the currently employed cesium beam clocks. A superior performance can, in particular, be noted for the GPS IIF RAFS and the Galileo PHM, which exhibit stability values close to the observational limit over a wide range of timescales. While the Allan deviation follows the expected linear trend in the double-logarithmic representation for most clocks and satellites, bumps of varying amplitude can frequently be recognized at timescales near one half of the orbital period. These may be caused by thermal variations in the onboard equipment as discussed above, but can also be related to radial orbit errors induced by, for example, an incomplete modeling of solar radiation pressure forces [5.95]. On the other hand, ADEV bumps observed at timescales less than 1000 s for some types of GNSS satellites can often be attributed to the function of the respective timekeeping system [5.78].

5.4 Relativistic Effects on Clocks

The technologies of clocks, timekeeping, and GNSS have long advanced to a state where the precision of the measurements is on the order of nanoseconds and where such performance is necessary for satellite-based global navigation. To achieve this level of precision and accuracy, corrections imposed by relativity must be taken into account in addition to errors in the measurement systems, instrumental errors on board a satellite and atmospheric propagation delays. The development of GPS first brought a system into application, in which the principles of the special and general theories of relativity are not merely a matter of scientific interest, but have become an engineering necessity.

To understand the relationship between Coordinated Universal Time (UTC) reference timescale, its dependence on the International Atomic Time (TAI), their relation to the rotation of the Earth, as well as the time maintained and used by individual GNSSs, the relativistic relationship of timekeeping in and around the Earth is necessary [5.96]. In a relativistic formulation it is also necessary to have a clear understanding of the relationship of space-time reference systems in that framework. These relationships have been defined primarily by resolutions of the various international scientific organizations. The most important of these resolutions are:

1. International Astronomical Union (IAU) Resolution A4 (1991) defining the Geocentric Celestial Reference System (GCRS), the Barycentric Celestial Reference System (BCRS) and their time coordinates. IAU Resolution B1 (2000) further refines the BCRS definition.
2. International Union of Geodesy and Geophysics (IUGG) Resolution 2 (2007; see [5.96, Annex C]) defining the Geocentric Terrestrial Reference System (GTRS), along with the International Terrestrial Reference System (ITRS).

The nomenclature used here follows the IAU/IUGG framework in that the GCRS is known as the Earth-Centered Inertial (ECI) coordinate system, the GTRS (in practice, the ITRS) is known as the Earth-Centered Earth-Fixed (ECEF) coordinate system, and the BCRS is the barycentric coordinate system.

5.4.1 Relativistic Terms

GPS provided the initial theoretical model for relativistic synchronization and time comparison for GNSS development and operation as well as a laboratory for the validation of relativistic algorithms. However, the effects of relativity on precise orbiting vice orbit clocks

and the systems operating them should be understood in the context of global timekeeping. Relativistic formalism and relationships are based on observations and measurements between different frames of reference. Consequently an understanding of the different frames of reference and how they are realized in actual use is needed. This section will discuss a description of the relativistic effects on orbiting and Earth bound clocks and the implications for GNSS and other satellite-based time and position determination systems. This discussion is a prelude to understanding the issues of global timekeeping and GNSS timekeeping systems which are directly analogous.

GNSS internal timescales provide a basis for maintaining system internal synchronization and precision measurements between the elements of the system. The GNSS time is based on time maintained by sets of atomic clocks that are effectively generating versions of atomic time. They are linked to timekeeping centers so they can provide the most versatile and effective means of disseminating precise global time available.

The definition of some key relativistic terms and their meaning follows:

Proper time τ_p is the actual reading of a clock or the local time in the clock's own frame of reference.

Coordinate time t is the independent variable in the equations of motion of material bodies and in the equations of propagation of electromagnetic waves. It is a mathematical coordinate in the four-dimensional space-time of the coordinate system. For a given event, the coordinate time has the same value everywhere. Coordinate times are not measured; rather, they are computed from the proper times of clocks.

Space-time interval. The relation between coordinate time and proper time depends on the clock's position and state of motion in its gravitational environment and is derived by integration of the space-time interval. In the comparison of the proper times of two clocks, the coordinate time is ultimately eliminated. Thus the relativistic transfer of time between clocks is independent of the coordinate system. The coordinate system may be chosen arbitrarily on the basis of convenience.

In general, the space-time interval ds is described by

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu = g_{00} c^2 dt^2 + 2g_{0j} c dt dx^j + g_{ij} dx^i dx^j, \quad (5.27)$$

where $g_{\mu\nu}$ are the components of the metric. In the notation used here a Greek index assumes the range 0, 1, 2, 3 and a Latin index assumes the range 1, 2,

3. A repeated index implies summation on that index. The metric depends upon the gravitational potentials, angular velocity and linear acceleration of the reference frame. Upon a transformation of the coordinates, the space-time interval remains invariant. Therefore the metric $g_{\mu\nu}$ transforms as a second-order covariant tensor.

The general expression for the relationship between proper time τ_p and the coordinates of the chosen coordinate system, comprising the coordinate time $x^0 \equiv ct$ and the spatial coordinates x^j , is given by

$$ds^2 = g_{00} c^2 dt^2 + 2g_{0j} c dt dx^j + g_{ij} dx^i dx^j = -c^2 d\tau_p^2. \quad (5.28)$$

Therefore $dt = d\tau_p$ for a clock at rest in an inertial frame of reference, for which $dx^j = 0$ and $-g_{00} = 1$, $g_{0j} = 0$, and $g_{ij} = \delta_{ij}$. The elapsed coordinate time corresponding to the measured proper time as registered by a clock along a path between points A and B is

$$\Delta t = \pm \int_A^B \frac{1}{\sqrt{-g_{00}}} \sqrt{1 + \frac{1}{c^2} \left(g_{ij} + \frac{g_{0i} g_{0j}}{-g_{00}} \right) \frac{dx^i}{d\tau_p} \frac{dx^j}{d\tau_p}} d\tau_p + \frac{1}{c} \int_A^B \frac{g_{0j}}{-g_{00}} \frac{dx^j}{d\tau_p} d\tau_p. \quad (5.29)$$

For an electromagnetic signal, the space-time interval is

$$ds^2 = g_{00} c^2 dt^2 + 2g_{0j} c dt dx^j + g_{ij} dx^i dx^j = 0. \quad (5.30)$$

The speed of light is c in every inertial frame of reference. The elapsed coordinate time of propagation along a path between points A and B is

$$\Delta t = \pm \int_A^B \frac{1}{\sqrt{-g_{00}}} \sqrt{1 + \frac{1}{c^2} \left(g_{ij} + \frac{g_{0i} g_{0j}}{-g_{00}} \right) dx^i dx^j} + \frac{1}{c} \int_A^B \frac{g_{0j}}{-g_{00}} dx^j, \quad (5.31)$$

where the expression in parenthesis

$$\gamma_{ij} \equiv g_{ij} + \frac{g_{0i} g_{0j}}{-g_{00}}$$

represents the metric of three-dimensional space and

$$d\rho = \sqrt{\gamma_{ij} dx^i dx^j}$$

represents the increment of three-dimensional distance.

5.4.2 Coordinate Timescales

For practical purposes, different types of coordinate times are distinguished:

Geocentric Coordinate Time (TCG) is the coordinate time in a coordinate system with origin at the Earth's center (ECI or ECEF).

Terrestrial Time (TT) is the coordinate time that is rescaled from TCG so that it has approximately the same rate as the proper time of a clock at rest on the geoid. The geoid is the surface of constant gravity potential, which is closely approximated by mean sea level. The relationship between TCG and TT is defined such that

$$\frac{d(TT)}{d(TCG)} \equiv 1 - L_G ,$$

where

$$L_G \equiv 6.969290134 \cdot 10^{-10} \approx 60.2 \mu\text{s/d}$$

as discussed below. The value of L_G is a defined constant. Consequently,

$$\begin{aligned} TCG - TT &= L_G(TCG - TCG_0) \\ &= \frac{L_G}{1 - L_G}(TT - TT_0) , \end{aligned}$$

where TCG_0 and TT_0 correspond to JD 2 443 144.5 TAI (1977 Jan. 1, 0 h). The practical realization of TT is

$$TT = TAI + 32.184 \text{ s} . \quad (5.32)$$

Barycentric Coordinate Time (TCB) is the coordinate time in a coordinate system with origin at the solar system barycenter. The coordinate time difference between TCB and TCG is a transformation that depends on both time and position. This timescale is important to be used when the satellite is outside Earth orbit and the influence of the Solar System needs to be taken into account.

5.4.3 Geocentric Coordinate Systems

The following section will discuss the transformation between the proper time of an ideal clock (one that exactly realizes the SI second) and coordinate time in a geocentric coordinate systems.

Earth-Centered Inertial Coordinate System

The coordinate time associated with an Earth-centered inertial (ECI) coordinate system is TCG. Through terms

of order $1/c^2$, the components of the metric tensor in this coordinate system are

$$\begin{aligned} -g_{00} &= 1 - \frac{2U}{c^2} , \\ g_{0j} &= 0 , \\ g_{ij} &= \left(1 + \frac{2U}{c^2}\right) \delta_{ij} , \end{aligned} \quad (5.33)$$

where U is the gravitational potential. The elapsed TCG in the ECI coordinate system corresponding to the elapsed proper time as registered by a clock moving along a path between points A and B with velocity v is given by

$$\Delta t = \int_A^B \left(1 + \frac{1}{c^2}U + \frac{1}{2c^2}v^2\right) d\tau_p . \quad (5.34)$$

The Earth's potential U at radial distance r , geocentric latitude ϕ , and longitude λ may be expressed as an expansion in spherical harmonics

$$\begin{aligned} U(r, \phi, \lambda) &= \frac{GM_\oplus}{r} \left[1 + \sum_{n=2}^{\infty} \sum_{m=0}^n \left(\frac{R_\oplus}{r}\right)^n \right. \\ &\quad \left. \times P_{nm}(\sin \phi)(C_{nm} \cos m\lambda + S_{nm} \sin m\lambda) \right] \end{aligned} \quad (5.35)$$

with coefficients C_{nm} and S_{nm} . Here, GM_\oplus is the gravitational coefficient of the Earth and R_\oplus its equatorial radius. Furthermore, the factors P_{nm} denote the associated Legendre functions of degree n and order m .

For practical applications it is sufficient to include only the Earth oblateness correction and approximate the gravitational potential as

$$\begin{aligned} U &= \frac{GM_\oplus}{r} \\ &\quad + J_2 \frac{GM_\oplus}{r} \left(\frac{R_\oplus}{r}\right)^2 \frac{1}{2} (1 - 3 \sin^2 \phi) , \end{aligned} \quad (5.36)$$

where $J_2 = -C_{2,0} \approx 1.08 \cdot 10^{-3}$ is the leading zonal gravity coefficient.

Even for a clock at rest on the surface of the rotating Earth, it is necessary to account for its velocity $\mathbf{v} = \boldsymbol{\omega} \times \mathbf{r}$ in the ECI coordinate system, where $\boldsymbol{\omega}$ is the angular velocity of the Earth and \mathbf{r} is the position of the clock. Thus TCG elapsed as the clock records proper time $\Delta\tau_p$

is

$$\begin{aligned}\Delta t &= \int_A^B \left(1 + \frac{1}{c^2} U + \frac{1}{2c^2} \|\boldsymbol{\omega} \times \mathbf{r}\|^2 \right) \\ &= \int_A^B \left(1 + \frac{1}{c^2} W \right) d\tau_p, \end{aligned} \quad (5.37)$$

where

$$\begin{aligned}W &= U + \frac{1}{2} \|\boldsymbol{\omega} \times \mathbf{r}\|^2 \\ &= U + \frac{1}{2} \omega^2 r^2 \cos^2 \phi \end{aligned} \quad (5.38)$$

is the gravity potential.

As the gravity potential W_0 over the surface of the geoid is constant, it may be evaluated on the equator and is approximately given by

$$W_0 \approx \frac{GM_\oplus}{R_\oplus} \left(1 + \frac{1}{2} J_2 \right) + \frac{1}{2} \omega^2 R_\oplus^2. \quad (5.39)$$

The current best estimate of W_0 is $6.2636856 \cdot 10^7 \text{ m}^2/\text{s}^2$.

According to (5.37), the TCG in the ECI coordinate system that corresponds to the proper time $\Delta\tau_{p0}$ measured by a clock at rest on the geoid is

$$\begin{aligned}\Delta t \equiv \text{TCG} &= \left(1 + \frac{W_0}{c^2} \right) \Delta\tau_{p0} \\ &\approx (1 + L_G) \Delta\tau_{p0}, \end{aligned} \quad (5.40)$$

where $L_G \equiv 6.969290134 \cdot 10^{-10}$. By convention, the value of L_G is a defined constant. It represents the best available value of W_0/c^2 at the time of its definition in 2000 [5.96].

TT is obtained by rescaling TCG by the factor $1 - L_G$. Thus

$$\Delta t' \equiv \text{TT} = (1 - L_G) \text{TCG}. \quad (5.41)$$

It follows that

$$\text{TT} = (1 - L_G)(1 + L_G) \Delta\tau_{p0} \approx \Delta\tau_{p0}$$

to within a few parts in 10^{18} .

For a clock on an Earth-orbiting satellite, the orbit may be regarded as Keplerian (unperturbed) in the first approximation (Chap. 3). The potential at distance r from the Earth's center is approximately $U = GM_\oplus/r$. Therefore the increment of TCG is

$$\Delta t = \int_A^B \left(1 + \frac{1}{c^2} \frac{GM_\oplus}{r} + \frac{1}{2c^2} v^2 \right) d\tau_p. \quad (5.42)$$

The variation of the satellite velocity v with distance r is determined from the conservation of the specific energy

$$\varepsilon = \frac{1}{2} v^2 - U = \frac{1}{2} v^2 - \frac{GM_\oplus}{r}, \quad (5.43)$$

which amounts to

$$\varepsilon = -\frac{GM_\oplus}{2a} \quad (5.44)$$

for an orbit of semimajor axis a . Therefore, to this order of approximation, the elapsed coordinate time is

$$\begin{aligned}\Delta t &= \int_A^B \left(1 - \frac{1}{c^2} \frac{GM_\oplus}{2a} + \frac{1}{c^2} \frac{2GM_\oplus}{r} \right) d\tau_p \\ &= \left(1 - \frac{1}{c^2} \frac{GM_\oplus}{2a} \right) \Delta\tau_p \\ &\quad + \frac{2GM_\oplus}{c^2} \int_{t_0}^t \frac{1}{r} dt. \end{aligned} \quad (5.45)$$

In the last integral $d\tau_p$ has been replaced by dt as this term is a relativistic correction of order $1/c^2$.

For a Keplerian orbit the radial distance is given by

$$r = a(1 - e \cos E),$$

where e is the orbital eccentricity and E is the eccentric anomaly. The eccentric anomaly is determined from the mean anomaly by Kepler's equation

$$M \equiv n\Delta t = E - e \sin E,$$

where

$$n \equiv \frac{2\pi}{T} = \sqrt{\frac{GM_\oplus}{a^3}}$$

is the mean motion and T is the orbital period (Chap. 3). Therefore, the TCG elapsed as the clock records proper time $\Delta\tau_p$ is approximately

$$\begin{aligned}\Delta t &= \int_A^B \left(1 - \frac{1}{c^2} \frac{GM_\oplus}{2a} + \frac{1}{c^2} \frac{2GM_\oplus}{r} \right) d\tau_p \\ &= \left(1 + \frac{3}{2} \frac{1}{c^2} \frac{GM_\oplus}{a} \right) \Delta\tau_p \\ &\quad + \frac{2}{c^2} \sqrt{GM_\oplus a} \cdot e \sin E. \end{aligned}$$

The second term is a periodic correction due to the orbital eccentricity that causes a residual variation in distance and velocity given by

$$\Delta t_{\text{ecc}} = \frac{2}{c^2} \sqrt{GM_{\oplus} a} \cdot e \sin E = \frac{2}{c^2} (\mathbf{v} \cdot \mathbf{r}) . \quad (5.46)$$

To compare the proper time of a clock on a satellite with the proper time of a clock at rest on the geoid, it is necessary to convert from TCG to TT. By (5.41) and (5.42), the result is (TT)

$$\begin{aligned} \Delta t' &= (1 - L_G) \Delta t \\ &= \int_A^B \left(1 + \frac{1}{c^2} (U - W_0) + \frac{1}{2} \frac{1}{c^2} v^2 \right) d\tau_p . \end{aligned}$$

Since $\Delta t' \approx \Delta \tau_{p0}$, the interval of proper time recorded by a clock at rest on the geoid, which corresponds to the interval of proper time recorded by a clock on the satellite, is therefore given by

$$\begin{aligned} \Delta \tau_{p0} &= \left(1 + \frac{3}{2} \frac{1}{c^2} \frac{GM_{\oplus}}{a} - \frac{1}{c^2} W_0 \right) \Delta \tau_p \\ &\quad + \frac{2}{c^2} \sqrt{GM_{\oplus} a} \cdot e \sin E . \end{aligned} \quad (5.47)$$

It comprises a rate difference at the level of few parts in 10^{10} for common GNSS satellites as well as the eccentricity dependent periodic part with representative amplitudes of 10–100 ns. Both contributions and their practical implications for navigation satellite systems are further discussed in Sect. 5.4.5.

At the subnanosecond level of precision, it is necessary to take into account the orbital perturbations due to the harmonics of the Earth's gravitational potential, the tidal effects of the Moon and the Sun, as well as solar radiation pressure. Leading contributions result from the J_2 perturbation of the satellite position and velocity. Following [5.97], these give rise to a supplementary correction

$$\begin{aligned} \delta \Delta \tau_{p0} &= \frac{7}{2} \frac{GM_{\oplus} R_{\oplus}^2}{a^3 c^2} J_2 \left(1 - \frac{3}{2} \sin^2 i \right) \Delta \tau_p \\ &\quad - \frac{3}{2} \frac{R_{\oplus}^2}{a^2 c^2} J_2 \sqrt{GM_{\oplus} a} \sin^2 i \sin 2u \end{aligned} \quad (5.48)$$

that needs to be considered on top of (5.47). Here, i denotes the inclination of the satellite orbit while u is the argument of latitude. Besides a drift correction, the J_2 contribution comprises a harmonic term of about 0.1 ns amplitude with a twice-per-revolution periodicity.

To fully account for the J_2 perturbation in the potential of (5.36), it is necessary to perform a numerical

integration of the orbit and a numerical integration of (5.42). The tidal effects of the Moon and the Sun and solar radiation pressure should also be considered. For low-Earth orbits, both the zonal and tesseral gravitational harmonics are important and the usual eccentricity correction of (5.46) is no longer accurate. In this case, it is likewise preferable to integrate the orbit and integrate (5.42) numerically including the higher-order harmonics of the Earth's gravitational potential.

Earth-Centered Earth-Fixed Coordinate System

Through terms of order $1/c^2$, the metric tensor components in the rotating Earth-centered Earth-fixed (ECEF) coordinate system are given by

$$\begin{aligned} -g_{00} &= 1 - \frac{2U}{c^2} - \frac{||\boldsymbol{\omega} \times \mathbf{r}||^2}{c^2} \\ &= 1 - \frac{2W}{c^2} \\ g_{0j} &= \frac{(\boldsymbol{\omega} \times \mathbf{r})_j}{c} \\ g_{ij} &= \delta_{ij} . \end{aligned} \quad (5.49)$$

Using coordinate time TT, the elapsed coordinate time is

$$\begin{aligned} \Delta t' &= \int_A^B \left(1 - \frac{1}{c^2} gh + \frac{1}{2} \frac{1}{c^2} (v')^2 \right) d\tau_p \\ &\quad + \frac{1}{c^2} \int_A^B (\boldsymbol{\omega} \times \mathbf{r}) \cdot \mathbf{v}' d\tau_p , \end{aligned} \quad (5.50)$$

where h is the height of the clock above the geoid, g is the local acceleration of gravity, v' is the velocity of the clock relative to the geoid, and \mathbf{r} and \mathbf{v}' are the position and velocity vectors of the clock in the ECEF frame. It is assumed that h is small. For high accuracy, the variation of g with latitude and elevation should also be taken into account.

The second integral of (5.50) is the Sagnac effect for a transported clock. This effect may be expressed as

$$\begin{aligned} \Delta t_{\text{Sagnac}} &= \frac{1}{c^2} \int_A^B (\boldsymbol{\omega} \times \mathbf{r}) \cdot \mathbf{v}' d\tau_p \\ &= \frac{1}{c^2} \int_A^B (\omega R_{\oplus} \cos \phi) (v' \cos \theta) d\tau_p \\ &= \frac{\omega R_{\oplus}^2}{c^2} \int_A^B \cos^2 \phi d\lambda = \frac{2\omega A}{c^2} , \end{aligned} \quad (5.51)$$

where ϕ is the latitude, λ is the longitude, $v' \cos \theta$ is the eastward component of the velocity, and A is the projection onto the equatorial plane of the area swept out by the position vector with respect to the center of the Earth (positive for the eastward direction and negative for the westward direction).

5.4.4 Propagation of Signals

This section deals with the computation of the coordinate time of propagation of signals when the transmitter and receiver positions are both given as expressed in the ECI, ECEF, and barycentric coordinate systems.

These equations apply in all cases. In particular, they must be used when setting the parameters of clocks on satellites that are steered to clocks on the Earth.

Propagation in ECI Coordinate System

When considering computation in an ECI coordinate system, the coordinate time of propagation (TCG) may be considered as the sum of a geometric part and a gravitational part. The geometric part is

$$\Delta t \approx \frac{1}{c} \int_{\text{path}} \sqrt{g_{ij} dx^i dx^j} = \frac{\rho}{c}, \quad (5.52)$$

where $g_{ij} \approx \delta_{ij}$ and ρ is the geometric path length of the signal path.

If the signal is transmitted at coordinate time t_T and is received at coordinate time t_R , the TCG of propagation over the path is

$$\begin{aligned} \Delta t &= \frac{\rho}{c} = \frac{1}{c} |\mathbf{r}_R(t_R) - \mathbf{r}_T(t_T)| \\ &\approx \frac{1}{c} |\Delta \mathbf{r} + \mathbf{v}_R(t_R - t_T)| \\ &\approx \frac{1}{c} |\Delta \mathbf{r}| + \frac{1}{c^2} \Delta(\mathbf{r} \cdot \mathbf{v}_R), \end{aligned} \quad (5.53)$$

where the transmitter has position \mathbf{r}_T and the receiver has position \mathbf{r}_R and velocity \mathbf{v}_R and where $\Delta \mathbf{r} \equiv \mathbf{r}_R(t_T) - \mathbf{r}_T(t_T)$ is the difference between the position of the receiver and the transmitter at the coordinate time of transmission t_T . The correction to the coordinate time due to the receiver velocity is

$$\Delta t_{\text{vel}} \approx \frac{\Delta \mathbf{r} \cdot \mathbf{v}_R}{c^2}. \quad (5.54)$$

Note that additional terms of order $1/c^3$ may amount to several picoseconds, depending on the configuration.

To consider the effect of the gravitational potential on an electromagnetic signal, it is necessary to include

the potential in both the spatial and temporal parts of the metric. The components of the metric are

$$\begin{aligned} -g_{00} &= 1 - \frac{2U}{c^2}, \\ g_{0j} &= 0, \\ g_{ij} &= \left(1 + \frac{2U}{c^2}\right) \delta_{ij}. \end{aligned} \quad (5.55)$$

Therefore, the elapsed TCG is

$$\begin{aligned} \Delta t &\approx \frac{1}{c} \int_{\text{path}} \sqrt{\frac{g_{ij}}{-g_{00}}} dx^i dx^j \\ &\approx \frac{1}{c} \left(1 + \frac{2U}{c^2}\right) \sqrt{\delta_{ij} dx^i dx^j} \\ &= \frac{\rho}{c} + \frac{1}{c^3} \int_{\text{path}} 2U d\rho. \end{aligned} \quad (5.56)$$

The gravitational time delay is

$$\Delta t_{\text{delay}} = \frac{2GM_{\oplus}}{c^3} \ln \left(\frac{R+r+\rho}{R+r-\rho} \right), \quad (5.57)$$

where R and r are the distances from the geocenter to the transmitter and receiver, respectively.

The gravitational delay typically amounts to a few tens of picoseconds for a path between a satellite and Earth. The total TCG is the sum of the terms in equations (5.53) and (5.57).

The coordinate time of propagation (TT) is

$$\begin{aligned} \Delta t' &= (1 - L_G) \Delta t \\ &= \frac{\rho}{c} - L_G \frac{\rho}{c} + \frac{2GM_{\oplus}}{c^3} \ln \left(\frac{R+r+\rho}{R+r-\rho} \right). \end{aligned} \quad (5.58)$$

This is the time interval that would be measured by a clock on the geoid. For example, a signal sent from a geostationary satellite with orbital radius 42 164 km to a clock on the equator at the same longitude would have a path delay of -27 ps. For a GPS satellite at an elevation angle of 40° , the second and third terms nearly cancel so that the path delay is -3 ps.

Propagation in ECEF Coordinate System

When considering signal propagation in an ECEF coordinate system, the geometric part of the TCG is

$$\Delta t = \frac{1}{c} \int_{\text{path}} \sqrt{g_{ij} dx^i dx^j} + \frac{1}{c} \int_{\text{path}} g_{0j} dx^j. \quad (5.59)$$

The metric components are

$$\begin{aligned} -g_{00} &\approx 1, \\ g_{0j} &= \frac{(\boldsymbol{\omega} \times \mathbf{r})_j}{c}, \\ g_{ij} &\approx \delta_{ij}, \end{aligned} \quad (5.60)$$

where \mathbf{r} is the position vector of a point on the signal path. The coordinate time (TT) is $\Delta t' = (1 - L_G)\Delta t$.

The first term of (5.59) is ρ'/c , where ρ' is the Euclidean path length in the ECEF coordinate system. If the transmitter has position \mathbf{r}_T and the receiver has position \mathbf{r}_R and velocity \mathbf{v}'_R , then

$$\begin{aligned} \frac{\rho'}{c} &= \frac{1}{c} |\mathbf{r}_R(t_R) - \mathbf{r}_T(t_T)| \\ &\approx \frac{1}{c} |\Delta \mathbf{r} + \mathbf{v}'_R(t_R - t_T)| \\ &\approx \frac{1}{c} |\Delta \mathbf{r}| + \frac{1}{c^2} \Delta \mathbf{r} \cdot \mathbf{v}'_R, \end{aligned} \quad (5.61)$$

where $\Delta \mathbf{r} \equiv \mathbf{r}_R(t_T) - \mathbf{r}_T(t_T)$.

The second term of (5.59) is the Sagnac effect. Therefore,

$$\begin{aligned} \Delta t_{\text{Sagnac}} &= \frac{1}{c^2} \int_A^B (\boldsymbol{\omega} \times \mathbf{r}) \cdot \mathbf{v}' d\tau_p \\ &= \frac{1}{c^2} \int_A^B (\boldsymbol{\omega} \times \mathbf{r}) \cdot d\mathbf{r} \\ &= \frac{1}{c^2} \int_A^B \boldsymbol{\omega} \cdot (\mathbf{r} \times d\mathbf{r}) \\ &= 2 \frac{1}{c^2} \int_A^B \boldsymbol{\omega} \cdot d\mathbf{A} = \frac{2\omega A}{c^2}, \end{aligned} \quad (5.62)$$

where A is the projection onto the equatorial plane of the area formed by the center of rotation and the endpoints of the signal path. The gravitational delay must also be considered to compute the total time of propagation.

5.4.5 Relativistic Offset for GNSS Satellite Clocks

As noted before, the Global Positioning System as well as all other GNSSs must incorporate relativistic effects in their normal operation. They provide a means of performing time and position measurements with satellite

clocks on a practical scale and a validation of the accuracy and consistency of the algorithms is used in such measurements across a broad area of applications.

For measurements with a precision at the nanosecond level, there are three relativistic effects that must be taken into account. First, there is the effect of time dilation. The velocity of a moving clock causes it to appear to run slow relative to a clock on the Earth. GPS satellites revolve around the Earth with an orbital period of 11.967 h and a velocity of 3.874 km/s. Thus, on account of its velocity, a GPS satellite clock appears to run slow by 7 $\mu\text{s/d}$. Second, there is the effect of gravitational redshift. At an altitude of 20 184 km the difference in gravitational potential causes the satellite clock to appear to run fast by 45 $\mu\text{s/d}$. In addition, the effects contributed by the velocity of rotation and gravitational potential of the rotating geoid must be included. The net effect of time dilation and gravitational redshift is that the satellite clock appears to run fast by approximately 38 $\mu\text{s/d}$ when compared to a similar clock on the Earth's surface, which is an enormous rate difference for a clock with a precision of a few nanoseconds. To compensate for this large secular effect, the GPS clock is given a fractional rate offset prior to launch of $-4.465 \cdot 10^{-10}$ from its nominal frequency of exactly 10.23 MHz, so that on average it appears to run at the same rate as a clock on the ground. The actual frequency of the satellite clock prior to launch is thus 10.22999999543 MHz. Similar considerations apply for the other navigation satellite systems, even though the apparent frequency is different for each constellation due to the specific orbital altitude and velocity of the various satellites.

Although the GPS orbits are nominally circular, there is always some residual eccentricity. The eccentricity causes the orbit to be slightly elliptical. Thus the velocity and gravitational potential vary slightly over one revolution and, although the principal secular effect is compensated by a rate offset, there remains a small residual variation that is proportional to the eccentricity. For example, with an orbital eccentricity of 0.02, there is a relativistic sinusoidal variation in the apparent clock time having an amplitude of 46 ns at the orbital period. By convention [5.98], the eccentricity-dependent periodic relativistic effect is removed in precise clock products of all GNSSs produced by the IGS and other providers, to obtain an essentially linear variation of the reported clock offset. With the exception of GLONASS [5.99], the same holds for broadcast clock offset values transmitted by the various GNSSs as part of their navigation messages [5.100–104]. This correction must therefore be calculated and taken into account in the user's receiver.

The third relativistic effect is associated with the universality of the speed of light. Thus the displacement of the receiver relative to an inertial frame during the time of flight of the signal must be included. In the Earth's rotating frame of reference, this property is called the Earth rotation correction or Sagnac effect. For a receiver at rest on the rotating geoid observing a GPS satellite, the maximum correction is 133 ns.

GPS has served as a laboratory for doing physics at the one-to-ten nanosecond level. The consistent application of relativity to GPS Time and position measurements has been demonstrated by the operational precision of the system and by numerous experiments designed to test these individual effects over a wide range of conditions.

A more robust treatment of relativity will be required for clock modeling and orbit determination in future GNSS generations that are looking for greater ac-

curacy. At the subnanosecond level, relativistic effects must be included that are not modeled in the present system. One of the most important comes from the contribution to the redshift from the gravitational potential harmonic due to Earth oblateness [5.97]. There is a secular effect of approximately 0.5 ns/day and a periodic effect at half the orbit period with amplitude of 0.04 ns (corresponding to about 1 cm). While partly masked by thermal clock or bias variations, this effect has become discernible with the high precision provided by modern rubidium or hydrogen maser clocks of new GNSS satellites [5.88]. At the few picosecond level, it is also necessary to consider the tidal potentials of the Sun and Moon and the Earth's gravitation has an effect on the speed of propagation of light itself. That is, the gravitational potential causes the speed of propagation to depart slightly from the value of c .

5.5 International Timescales

The development of atomic clocks and their use in GNSS has made precise and accurate clock measurements and comparisons available globally. One of the first applications of this technology was in the comparison and generation of global timescales. Since these clocks generate *atomic time*, their use in GNSS is linked to global timekeeping of atomic time and the accuracy of this timekeeping is now such that the principles of relativity need to be taken into account.

Atomic time has become the basis of all physically realized timescales. A version of atomic time has been maintained continuously in various laboratories since 1955 although not formally adopted as an international timescale until 1971. Prior to the creation of the Bureau International de l'Heure (BIH) in 1920 at the Paris Observatory, timescales were based wholly upon astronomical observations and were not internationally agreed. The unit of time of the timescale in use was the second and it was likewise based on the astronomical observation of the length of a day.

With advent and operation of cesium atomic standards in the 1950s, and broadcast systems such as Long Range Navigation (LORAN) enabling accurate international comparison of these standards, these led to the initial form of Atomic Time (AT). The formation of International Atomic Time (TAI) was recommended by the International Astronomical Union (IAU) in 1967, the International Union of Radio Science (URSI) in 1969 and the International Radio Consultative Committee (CCIR) of the International Telecommunication Union (ITU) in 1970.

The 14th General Conference on Weights and Measures (CGPM) approved the establishment of TAI in 1971 as the relativistic coordinate time scale whose unit interval is the second of the International System of Units (SI) as realized on the rotating geoid. This established the SI second as the standard for the unit of time based on the hyperfine frequency of cesium 133 [5.105, 106].

5.5.1 International Atomic Time (TAI)

The International Atomic Time (TAI) is the metrologic timescale maintained by the Bureau des Poids et Mesure (BIPM) since timekeeping responsibility was transferred to the BIPM from the BIH in 1988 [5.24, 107]. It is established as the basis of atomic clock comparison data supplied to the BIPM by participating timekeeping centers and laboratories and is generated by a particular algorithm for treating the data known as ALGOS [5.108].

TAI is defined as a coordinate timescale in a geocentric reference frame with the SI second as the scale unit realized on the rotating geoid. The fact that TAI is a coordinate timescale was determined by the Consultative Committee for the Definition of the Second in 1980 and the necessary framework for evaluating the relativistic terms in the establishment of TAI have been presented in Sect. 5.4.

The accuracy of TAI is a primary consideration in maintaining the SI second and providing a reliable reference scale in the long term [5.109]. The optimization

of the long-term stability is done at the expense of short-term accessibility. The calculation of TAI uses data over an extended period. Clock-comparison data are sent to the BIPM every ten days on days with the modified Julian date (MJD) ending in 9. Blocks of data covering 60 days are used in the calculation of the scale.

A period of 60 days was chosen to place the effective integration time of the scale at the transition between the flicker floor and the random walk frequency modulation of cesium clocks. Stability would therefore not be improved by a longer integration time. The period of 60 days is enough to smooth out the noise contributed by the time links (GNSS and other comparison techniques) and the white frequency modulation noise of the clocks. The monthly BIPM Circular T then alternates between provisional, based on only 30 days of data, and the definite, based on the full 60 days of data.

The determination of TAI [5.110] is then performed in three steps:

1. Calculation (using a post-processing, iterative procedure) of an intermediate scale, known as Echelle Atomique Libre (EAL) or Free Atomic Scale, using the clock comparison data and ALGOS
2. The evaluation of the duration of the scale unit of EAL using data from primary frequency standards and an optimum filter
3. The production of TAI from EAL by applying, if necessary, a correction to the scale interval of EAL to give a value as close as possible to the SI second. Correcting of the scale unit is known as *steering* and is done infrequently.

At a time t , the free atomic time scale EAL is defined in terms of the readings $h_i(t)$ of the group of N clocks, H_i , contributed by the various timing centers, expressed as

$$\text{EAL}(t) = \frac{\sum_{i=1}^N p_i [h_i(t) + h'_i(t)]}{\sum_{i=1}^N p_i} . \quad (5.63)$$

Here, p_i is the statistical weight assigned to clock H_i and $h'_i(t)$ is a time correction designed to ensure time and frequency continuity of the scale when either the weighting of individual clocks or the total number of clocks is changed [5.108, 111].

This expression cannot be used directly because the measured quantities that provide the basic data cannot be the readings of individual clocks but are comparisons between pairs of clocks. Such is the nature of time-keeping and clock measurements. At time t , the slowly

varying differences $\zeta_{ij}(t)$ between clocks H_i and H_j are written as $\zeta_{ij}(t) = h_i(t) - h_j(t)$.

The output of EAL is N values of the differences $x_i(t)$ defined by $x_i(t) = \text{EAL}(t) - h_i(t)$, where x_i are the differences between the individual clocks and the time defined by EAL. The difference can then be expressed as $x_i(t) - x_j(t) = -\zeta_{ij}(t)$, and (5.63) can be transformed into

$$\sum_{i=1}^N p_i x_i(t) = \sum_{i=1}^N p_i h'_i(t) . \quad (5.64)$$

In practice comparative data from the nonredundant system of $N - 1$ time links between the timing centers is employed to solve for these last two expressions.

The weight assigned to each clock is calculated in such a way as to favor the long-term stability of the resulting scale. This also minimizes the annual fluctuations and frequency drift that the predominately commercial frequency standards maintained at the timing centers with respect to the primary timescale frequency standards. An important feature of the ALGOS algorithm is the fact that the evaluation of the weight of the clock, although based upon data covering a whole year, takes into account the 60 days of data for which EAL is being specifically computed. It is thus possible to judge clocks on their actual performance during the interval of time during which EAL is being established. It is also possible to take into account of any abnormal behavior observed in an individual clock by adjusting its weight, if necessary to zero. This has proved useful on many occasions.

The weight is normally based on the variance $\sigma_i^2(6, \tau)$ of mean rate with respect to EAL calculated over two-monthly samples [5.108]. This variance, instead of the usual pair variance, was chosen because it gives greater reduction in the weight of clocks showing a frequency drift. The weights are obtained directly from

$$p_i = \frac{1000}{\sigma_i^2(6, \tau)} , \quad (5.65)$$

with σ_i expressed in nanoseconds per day, provided that over the current period of 60 days no abnormal behavior is apparent. In the case of abnormal behavior a weight of zero is assigned. A maximum weight of 100 is assigned to the 15% or so of clocks having $\sigma_i(6, \tau) \leq 3.16$ ns/d. The maximum weight is chosen to ensure that the scale is heavily biased in favor of the best clocks without allowing any one to become predominant, no clock having a contribution greater than about 2%.

The time correction term

$$h'_i(t) = a_i(t_0) + B_{ip}(t)(t - t_0) \quad (5.66)$$

is made up of two components, where $a_i(t_0)$ is simply the time differences between the clock H_i and EAL at a time t_0 , which is the time at the beginning of the 60-day period, and $B_{ip}(t)$ is the predicted difference in rate between H_i and EAL for the period between t_0 and t . The rate of clock H_i , for example, is defined by

$$\text{rate} = \frac{a_i(t_0 - t) - a_i(t_0)}{t - t_0}. \quad (5.67)$$

The prediction of $B_{ip}(t)$ is obtained by a one-step linear prediction based upon the previous value. This is justified by the fact that the period of 60 days is such that the predominant clock noise is random walk for which the most probable estimate for the value over the next period is simply that over the immediate preceding period. Having established the best estimate of EAL the transformation to TAI is made by the decision whether or not the rate of EAL differs sufficiently from the rate of the best primary standards to warrant a correction or *steering*. From 1984 to 1989, no steering was necessary and thus during this period TAI was in fact simply EAL with a constant frequency offset. Since then, several frequency changes of five parts in 10^{15} each have been found necessary.

Finally, the output of these calculations is presented in the monthly *Circular T* published by the BIPM and distributed to participating timing centers. This circular is alternately provisional or definitive depending upon whether it is issued in the middle or at the end of a 60-day calculation period.

5.5.2 Coordinated Universal Time (UTC)

The initial form of atomic time attempted prior to 1972 was to keep close agreement to astronomical time by adjusting both the frequency offset and fractional step adjustments of time broadcasts of atomic time signals with the Earth's rotation. Close coupling to the Earth's rotation was considered necessary to aid celestial navigation, however that system of adjustment was very difficult to coordinate between broadcast stations and provide a uniform accurate reference time. The present UTC system was adapted so that an approximation to the epoch of universal time (UT1), a version of universal time (sometimes called mean solar time) that is determined by the transit time of stars corrected for seasonal variations, and the interval of the SI second would be provided by a single scale. The history and development of these timescales is discussed in [5.112].

The current definition of Coordinated Universal Time (UTC) was developed by the Consultative Committee for International Radiocommunication (CCIR; now known as International Telecommunications Union Radiocommunication sector, ITU-R) [5.110]. It was a compromise solution to the international timescale between using TAI and continuing international time based on the rotation of the Earth. It is a stepped atomic time scale based on the rate of TAI adjusted by the addition or deletion of integer seconds, known as leap seconds, to maintain the time within 0.9 s of Universal Time (UT1). UTC is used to coordinate the reference time kept by the timing centers. The specific definition is maintained by ITU-R Recommendation TF.460.6 [5.110, 113]. Since UTC was adopted its use has grown considerably within the radio and telecommunications community.

UTC is specifically defined as $\text{TAI} - \text{UTC} = n$, an integer number of seconds, and $|\text{UT1} - \text{UTC}| < 0.9$ s. In 2014, UTC is behind TAI by 35 seconds. The integer number of seconds' difference is adjusted by the use of *leap seconds* (either positive or negative) to maintain the relationship of UTC to UT1. For other uses dependent upon coordination with Earth orientation the difference can be further refined by the additional term DUT1 recommended to be broadcast for further adjustment of UTC. DUT1 is the *predicted* difference, $\text{UT1} - \text{UTC}$ in integral multiples of 0.1 s. A user of UT1 can then adjust the accuracy to < 0.1 s. The decision of when to insert a leap second is determined by the variable change in the rotation rate of the Earth. The International Earth Rotation and Reference Service (IERS) monitors the rotation rate of the Earth along with the other Earth orientation parameters including the predicted value of DUT1. They determine the need to adjust UTC and advise the BIPM when to insert or delete the second.

As the needs of GNSS operations, broadcasting and timekeeping services require the generation and transmission of a *real-time* or immediate timescale and UTC itself is based on the rate of TAI which is a post-processed metrologic timescale, a real-time version was needed. This real-time version also needed to be produced from the same clocks and oscillators that produce the timekeeping data for TAI. To provide a real-time timescale timing centers produce a real-time representation of UTC and identify it by using the nomenclature UTC(k) where k is the timing center. UTC without the qualifying parentheses and k nomenclature would then identify the timescale referenced as the final international value determined by the BIPM. The centers maintaining a realization of UTC relating to the major GNSS and their values of UTC(k) are shown in Fig. 5.32.

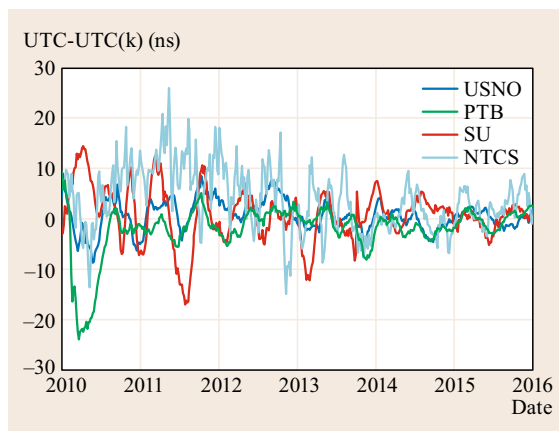


Fig. 5.32 Values of UTC(k) realizations of UTC maintained by associated GNSS timing centers (after [5.114])

These qualifications are made in ITU-R Recommendation TF.536 on Time-Scale Notation [5.115]. The final determination of UTC does not have a physical output and is available after a delay of two to four weeks in the form of an offset from the representation main-

tained by a participating laboratory. The value for both TAI and UTC are disseminated by the BIPM via the monthly publication of *Circular T*. The offset of TAI and UTC with respect to participating observatories or laboratories is given in terms of TAI or UTC minus UTC(k). For example, UTC (USNO) is the delivered real-time prediction of UTC as maintained by US Naval Observatory.

UTC is recognized as the basis of global time-keeping and telecommunications use and is the only timescale that is physically realized and disseminated. The CCIR in 1978 and the World Administrative Radio Conference (Geneva) in 1979 recommended that UTC should be used to designate the time in all international telecommunication activities. The ITU *Radio Regulations* define UTC as the timescale based on the SI second, specified in Recommendation ITU-R TF.460, and notes that UTC may be regarded as the general equivalent of mean solar time at the Greenwich meridian. The global reference for GNSS and satellite time and frequency applications is UTC. Its use in GNSS and managing with its discontinuous nature will be discussed in the following sections.

5.6 GNSS Timescales

In order to achieve synchronicity within a GNSS a stable common time reference must be globally available for the supporting ground segment to provide precise observations of the satellites for computation of the data products needed to operate the system and support the many users. This time reference is most commonly achieved in today's GNSSs through generation of its own real-time internal timescale in each system. Classically, reference time for a system was provided by establishing a specific master clock whose signals were transmitted or distributed as the reference time for all system measurements. For a system of global extent this master clock approach is difficult. It is virtually the same problem as global timekeeping but on a real-time basis.

The approach developed to address this problem is to generate an internal timescale in real time that is in effect a virtual system timescale made from the measurements of the clocks within the system. The system timescale is therefore not physically realized by an actual clock output within the system. This virtual time reference is possible since the GNSS must be supported by a global network of tracking sites to track and provide data to a centralized site in real time. Here, *real-time* means that as the observational data is collected the system data is incrementally generated for the system. The computation of the satellite ephemerides and other

navigation-related data, including the clock parameters of the satellites and ground supporting network, may be determined against a weighted average of the clock observed values.

The internal system reference timescale, GPS Time (GPST) for example, is produced in this manner. The process is similar to that of the international timescale being a weighted ensemble average of the physical clocks contributing to the operation of the system. Whereas the product of the international timescale is the publication of the offsets of the contributing time centers after the fact, a GNSS must have a real-time determination of the time within the system as the basis of system position and timing measurements. Consequently, the technique adopted has been the generation of a real-time weighed ensemble of the system's internal clocks that results in a time reference that is only physically realized by the clock output of the system's user receivers. The significant difference between the international timescale and the internal time reference is not just the real-time condition but the fact that the system time reference must deal with clocks of different characteristics. Classical ensembling for timescales deals with the same kind or similar clocks. With dissimilar types the combination of their stochastic characteristics has more complex computational difficulties.

An alternative approach to generating an internal system time reference is the more classical approach of considering the time reference to be independent of the system. Computationally the system's parameters would then consider time to be an independent value. The satellite and ground clocks could be monitored and maintained independently of the internal system operations through direct measurements of each clock, satellite or ground, using techniques such as two-way satellite time and frequency transfer (TWSTFT). This two-way technique is more commonly employed with communication satellites to compare clocks at ground sites but could be employed within a GNSS [5.116]. However, maintaining the clocks independently would lose the computational correlations with the other system parameters such as the satellite ephemerides since the precision and accuracy of the system measurements to the reference issue still needs to be minimized.

Consequently, within an internal time reference, observations are gathered passively and processed by the use of models, environmental measurements and other means of compensation of relevant observation effects necessary to relate the clocks to one another. A clock error model and timescale algorithm is then applied to the clock differences to produce estimates of the error of each clock relative to a stable virtual system timescale. Offset corrections of each satellite relative to its reference are then transmitted from each satellite in the form of predictions of the clock relative to System Time (ST) as part of its navigation message, along with the epoch of the prediction parameters so that offsets may be determined by the receiver software relative to the current epoch.

Variations of this approach are utilized by all of the major GNSSs, including GPS, GLONASS, Galileo, and BeiDou. Additional predicted corrections that relate the GNSS timescale to that of other systems and to UTC maintained by the timekeeping center associated with the particular system are also transmitted in the satellite system messages. Note that the GNSS ground segment upload interval for predicted navigation data to be trans-

mitted by the satellites is primarily determined by the intrinsic stability of the satellite clocks. The more stable the satellite clock the less frequent an upload of the predicted clock corrections is necessary. For example, GPS satellites navigation uploads and clock prediction corrections in the 2015 time period are nominally once per day.

Table 5.2 summarizes the current relationship between each of the main quartet of GNSS times and UTC as well as the system's strategy for maintaining time. All system times except GLONASS Time are continuous timescales that do not apply leap seconds as are currently applied to UTC.

GNSS receivers that are capable of simultaneously tracking satellites from multiple GNSSs are increasingly available. Because each system provides corrections relative to its own internal ST the user of multi-GNSS signals must either sacrifice the additional degrees of freedom necessary to estimate the offsets between each GNSS ST or, where available, utilize any cross-GNSS offsets that are may be included in the navigation messages.

In the case of GPST, each GPS satellite carries multiple atomic clocks, however only one is used to generate the satellite transmissions at a time. The satellite clocks and monitor station clocks contribute to the statistical formation of the continuous system time known as GPS Time [5.117, 118], which is specified to be within 1 μ s of UTC(USNO) modulo leap seconds since leap seconds are not inserted in GPS Time. GPST provides a time reference typically at a precision of better than 25 ns. However, depending upon the receiving equipment used in practice the precision can be 20 ns or better.

The epoch of GPST is midnight of January 5/6, 1980 UTC. Therefore, GPST is behind TAI by a constant value of 19 s. As of the beginning of 2016, GPST is ahead of UTC by 17 s which changes every time a leap second is applied to UTC. GPST is optimized for the navigation service, which requires short-term stability and uniform global distribution. For timing ap-

Table 5.2 GNSS Times versus UTC for the main quartet of GNSSs. Each offset is separated into the whole number of seconds and its subsecond component C_i . Furthermore, $n = \text{TAI} - \text{UTC}$ denotes the integer second offset between International Atomic Time and Coordinated Universal Time (e.g., $n = 36$ s starting on 1 July 2015)

UTC – GPST	$0 \text{ h} - n + 19 \text{ s} + C_0$	GPS Time (GPST) is steered to UTC(USNO), C_0 is required to be less than 1 μ s but is typically less than 20 ns
UTC – GLST	$-3 \text{ h} + 0 \text{ s} + C_1$	GLONASS Time (GLONASS Time) is steered to UTC(SU) including leap seconds. C_1 is required to be less than 1 ms. Note that GLONASS Time is offset from UTC by –3 hours corresponding to the offset of Moscow local time from the Greenwich meridian.
UTC – GST	$0 \text{ h} - n + 19 \text{ s} + C_2$	Galileo Time (GST) is steered to a set of European Union UTC(k) realization and C_2 is nominally less than 50 ns.
UTC – BDT	$0 \text{ h} - n + 33 \text{ s} + C_3$	BeiDou Time (BDT) is steered to UTC(NTSC) and C_3 is specified to be maintained less than 100 ns.

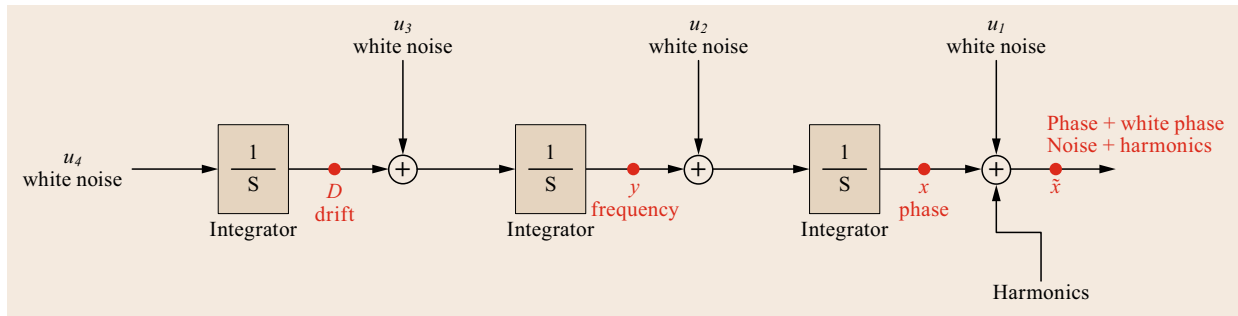


Fig. 5.33 IGS v2.0 Clock Model

plications that require a source of UTC, the real-time predicted offset of UTC (as realized at USNO) with respect to GPS Time is available from the GPS broadcast navigation message [5.100].

Like GPS, Galileo has adopted a uniform system time (GST) that does not contain leap seconds [5.101]. The starting epoch for GST is 00:00 on 22 August 1999 UTC (midnight between 21 and 22 August 1999). At this starting epoch GST was set to be ahead of UTC by 13 seconds so as to be consistent with GPST. The epoch for the Chinese BeiDou navigation satellite system time (BDT) is 00:00 on 1 January 2006 UTC and is related to UTC through UTC(NTSC) [5.102].

A primary requirement for a satellite navigation system is a uniformly precise system of time so that the navigation service is not interrupted by adjusting clocks. Such has not been the case with the Russian GLONASS satellite navigation system, which uses UTC(SU) offset by three hours as its system time [5.99].

The International GNSS Service (IGS; Chap. 33) established by the geodetic community also determines their own observational timescale, known as IGS Time (IGST) for coherency of the measure-

ment data collected globally by numerous agencies and sites. Its formation is similar to GPST [5.119, 120]. The clock model used in the second version of IGST implemented in 2011 [5.121] uses a basic four-state model for all the clocks, which is illustrated in Fig. 5.33.

The base model contains a deterministic phase x , the phase derivative (frequency) y , and the second derivative D of phase (drift) each with an independent random walk component. An additional phase state \tilde{x} is included to model a pure white phase noise and to couple it to any harmonic states. In accord with the observed clock behavior discussed in Sect. 5.3.6 four additional state parameters a_{ω_1} , b_{ω_1} , a_{ω_2} and b_{ω_2} are used to model up to two (once- and twice-per-revolution) harmonics that can appear in the satellite observations.

Acknowledgments. The authors would like to thank Dr. Joseph White, a colleague at the US Naval Research Laboratory, as well as Dr. Bob Nelson. Publication restrictions did not permit including Dr. White or Dr. Nelson in the authorship though many of the sections on individual atomic frequency standards and relativity were either contributed or edited by them.

References

- 5.1 D.B. Sullivan, D.W. Allan, D.A. Howe, F.L. Walls: *Characterization of Clocks and Oscillators*, TN-1337 (US National Institute of Standards and Technology, Gaithersburg 1990)
- 5.2 D.A. Howe, D.W. Allan, J.A. Barnes: Properties of signal sources and measurement methods, Proc. 35th Annu. Symp. Freq. Contr., Ft. Monmouth (IEEE, Piscataway 1981) pp. 669–716
- 5.3 J. Rutman, F.L. Walls: Characterization of frequency stability in precision frequency sources, Proc. IEEE **79**(7), 952–960 (1991)
- 5.4 T. Walter: Characterizing frequency stability: A continuous power-law model with discrete sampling, IEEE Trans. Instrum. Meas. **43**(1), 69–79 (1994)
- 5.5 W.J. Riley: *Handbook of Frequency Stability Analysis*, NIST Special Publication, Vol. 1065 (US National Institute of Standards and Technology, Gaithersburg 2008)
- 5.6 J.A. Barnes, A.R. Chi, L.S. Cutler, D.J. Healey, D.B. Leeson, T.E. McGunigal, J.A. Mullen Jr., W.L. Smith, R.L. Sydnor, R.F.C. Vessot, G.M.R. Winkler: Characterization of frequency stability, IEEE Trans. Instrum. Meas. **IM-20**(2), 105–120 (1971)
- 5.7 S.R. Stein: Frequency and time – Their measurement and characterization. In: *Precision Frequency Control*, Vol. 2, ed. by E.A. Gerber, A. Ballato (Academic Press, New York 1985) pp. 191–232

- 5.8 M.E. Frerking: Fifty years of progress in quartz crystal frequency standards, Proc. 50th IEEE Freq. Contr. Symp., Honolulu (1996) pp. 33–46
- 5.9 W.L. Smith: Quartz frequency standards and clocks – Frequency standards in general. In: *Precision Frequency Control*, Vol. 2, ed. by E.A. Gerber, A. Ballato (Academic Press, New York 1985) pp. 79–89
- 5.10 Hewlett-Packard: Fundamentals of Quartz Oscillators, Application Note 200–2 Electronic Counter Series (Hewlett-Packard Company, Palo Alto 1997)
- 5.11 C.S. Lam: A review of the recent development of MEMS and crystal oscillators and their impacts on the frequency control products industry, IEEE Ultrason. Symp., Beijing (IEEE, New Jersey 2008) pp. 694–704
- 5.12 J. Vanier, C. Audoin: Rubidium frequency standards. In: *The Quantum Physics of Atomic Frequency Standards*, Vol. 2, (Adam Hilger, Bristol 1989) pp. 1259–1350
- 5.13 S. Leschiutta: The definition of the ‘atomic’ second, *Metrologia* **42**, S10–S19 (2005)
- 5.14 *The International System of Units (SI)*, 8th edn. (Bureau International des Poids et Mesures, Paris 2006)
- 5.15 J. Vanier, C. Audoin: Cesium beam frequency standard, Part 1: Basic principles, frequency stability. In: *The Quantum Physics of Atomic Frequency Standards*, Vol. 2, ed. by EDITOR (Adam Hilger, Bristol 1989) pp. 613–781
- 5.16 A. Bauch: Caesium atomic clocks: Function, performance and applications, *Meas. Sci. Technol.* **14**(8), 1159–1173 (2003)
- 5.17 D. Kleppner, H.M. Goldenbergand, N.F. Ramsey: Theory of the hydrogen maser, *Phys. Rev.* **126**(2), 603–615 (1962)
- 5.18 D. Kleppner, H.C. Berg, S.B. Crampton, N.F. Ramsey, R.F.C. Vessot, H.E. Peters, J. Vanier: Hydrogen-maser principles and techniques, *Phys. Rev.* **138**(4A), A972–A983 (1965)
- 5.19 D.A. Howe, F.L. Walls, H.E. Bell, H. Hellwig: A small, passively operated hydrogen maser, Proc. 33rd Annu. Symp. Freq. Contr., Atlantic City (1979) pp. 554–562
- 5.20 H.T.M. Wang: An oscillating compact hydrogen maser, Proc. 34th Annu. Symp. Freq. Contr., Philadelphia (1980) pp. 364–369
- 5.21 W.M. Golding, R. Drullinger, A. De Marchi, W. Phillips: An atomic fountain frequency standard at NIST, Proc. 5th Symp. Freq. Stand. Metrol., Woods Hole, ed. by J.C. Bergquist (World Scientific, Singapore 1995) pp. 5–10
- 5.22 S.R. Jefferts, J. Shirley, T.E. Parker, T.P. Heavner, D.M. Meekhof, C. Nelson, F. Levi, G. Costanzo, A. De Marchi, R. Drullinger, L. Hollberg, W.D. Lee, F.L. Walls: Accuracy evaluation of NIST-F1, *Metrologia* **39**, 321–326 (2002)
- 5.23 R. Wynands, S. Weyers: Atomic fountain clocks, *Metrologia* **42**, s64–s79 (2005)
- 5.24 E.F. Arias: The metrology of time, *Phil. Trans. R. Soc. A* **363**, 2289–2305 (2005)
- 5.25 H.J. Metcalf, P. van de Straten: *Laser Cooling and Trapping* (Springer, New York 1999) pp. 156–164
- 5.26 M.A. Lombardi, T.P. Heavner, S.R. Jefferts: NIST primary frequency standards and the realization of the second, *NCSL Int. Meas.* **2**(4), 74–89 (2007)
- 5.27 S.L. Rolston, W.D. Phillips: Laser cooled neutral atom frequency standards, Proc. IEEE **79**(7), 943–951 (1991)
- 5.28 L. Cacciapiuoti, C. Salomon: Atomic clock ensemble in space, *J. Phys. Conf. Ser.* **327**(012049), 1–13 (2011)
- 5.29 J. Vanier, A. Godone, F. Levi, S. Micalizio: Atomic clocks based on coherent population trapping: Basic theoretical models and frequency stability, Proc. IEEE FCS 17th EFTF 2003, Tampa (2003) pp. 2–15
- 5.30 R. Lutwak, A. Rushed, M. Varghese, G. Tepolt, J. Lablanc, M. Mescher, D.K. Serkland, G.M. Peake: The miniature atomic clock – Preproduction results, Proc. IEEE FCS 21st EFTF 2007, Geneva (2007) pp. 1327–1333
- 5.31 F.-C. Chan, M. Joerger, S. Khanafseh, B. Pervan, O. Jakubov: Reducing the jitters – How a chip-scale atomic clock can help mitigate broadband interference, *GPS World* **5**(25), 44–51 (2014)
- 5.32 V. Giordano, S. Grop, B. Dubois, P.-Y. Bourgeois, Y. Kersalé, G. Haye, V. Dolgovskiy, N. Bucalovic, G. Di Domenico, S. Schilt, J. Chauvin, D. Valat, E. Rubiola: New generation of cryogenic sapphire microwave oscillators for space, metrology and scientific applications, *Rev. Sci. Instrum.* **83**(085113), 1–6 (2012)
- 5.33 E.A. Burt, S. Taghavi-Larigani, J.D. Prestage, R.L. Tjoelker: Compensated multi-pole mercury trapped ion frequency standard and stability evaluation of systematic effects, Proc. 7th Symp. Freq. Stand. Metrol., Pacific Grove, ed. by L. Maleki (World Scientific, Singapore 2009) pp. 321–328
- 5.34 N. Poli, C.W. Oates, P. Gill, G.M. Tino: Optical atomic clocks, *Riv. Nuovo Cimento* **36**(12), 555–624 (2013)
- 5.35 A.G. Smart: Optical-lattice clock sets new standard for timekeeping, *Phys. Today* **63**(3), 12–14 (2014)
- 5.36 J. Ye, H. Schnatz, L. Hollberg: Optical frequency combs: From frequency metrology to optical phase control, *IEEE J. Quantum Electron.* **9**(4), 1041–1058 (2003)
- 5.37 T.A. Stansell: The navy navigation satellite system: Description and status, *Navigation* **15**(3), 229–243 (1968)
- 5.38 B.W. Parkinson, T.A. Stansell, R.L. Beard, K. Gro-mov: A history of satellite navigation, *Navigation* **42**(1), 109–164 (1995)
- 5.39 R.L. Beard, J.A. Murray, J.D. White: GPS clock technology and navy PTI programs at the US Naval research laboratory, Proc. 18th Annu. PTI Meet., Reston (1987) pp. 37–53
- 5.40 A.B. Bassevich, P. Bogdanov, A.G. Gevorkyan, A. Tyulyakov: Glonass onboard time/frequency standards: Ten years of operation, Proc. 28th Annu. PTI Meet., Reston (1997) pp. 455–462
- 5.41 R. Fatkulin, V. Kossenko, S. Storozhev, V. Zvonar, V. Chebotarev: GLONASS space segment: Satellite constellation, GLONASS-M and GLONASS-K spacecraft, main features, ION GNSS 2012, Nashville (2012) pp. 3912–3930
- 5.42 P. Ro-chat, F. Droz, P. Mosset, G. Barmaverain, Q. Wang, D. Boving, L. Mattioni, M. Belloni,

- M. Gioia, U. Schmidt, T. Pike, F. Emma: The on-board galileo rubidium and passive maser, status and performance, Proc. IEEE FCS 2005, Vancouver (2005) pp. 26–32
- 5.43 J. Xie: Study to spaceborne rubidium atomic clocks characteristics and ground test requirements, Proc. CSNC 2014, Nanjing, Vol. III, ed. by J. Sun, W. Jiao, H. Wu, M. Lu (Springer, Berlin 2014) pp. 451–461
- 5.44 C. Li, T. Yang, L. Zhai, L. Ma: Development of new-generation space-borne rubidium clock, Proc. CSNC 2013, Wuhan, Vol. III, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin 2013) pp. 379–386
- 5.45 Y. Xie, P. Chen, S. Liu, T. Pei, Y. Shuai, C. Lin: Development of space mini passive hydrogen maser, Proc. CSNC 2015, Xi'an, Vol. III, ed. by J. Sun, J. Liu, S. Fan, X. Lu (Springer, Berlin 2015) pp. 343–349
- 5.46 N.D. Bhaskar, J. White, L. Mallette, T. McClelland, J. Hardy: A historical review of atomic frequency standards used in space systems, Proc. 50th IEEE FCS, Honolulu (1996) pp. 24–32
- 5.47 L. Mallette, J. White, P. Rochat: Space qualified frequency sources (clocks) for current and future GNSS applications, IEEE/ION PLANS 2010, Indian Wells (2010) pp. 903–908
- 5.48 J. White, R. Beard: Space clocks – Why they're different, Proc. 33rd PTI Meet., Long Beach, ed. by L.A. Breakiron (USNO, Washington, DC 2001) pp. 7–18
- 5.49 S. Nichols, J.D. White, F. Danzy: *Design and Ground Test of the NTS1 Frequency Standard System*, Naval Research Laboratory Report 7904, (Naval Research Laboratory, Washington 1975)
- 5.50 C.O. Alley, R. Williams, G. Singh, J. Mullendore: Performance of the new Ephraim optically pumped rubidium frequency standards and their possible application in space relativity experiments, Proc. 4th PTI Plan. Meet., Greenbelt (1972) pp. 29–40
- 5.51 M.J. Van Melle: Cesium and rubidium frequency standards status and performance on the GPS program, Proc. 27th Annu. PTI Meet., San Diego (1996) pp. 167–180
- 5.52 W.J. Riley: A rubidium clock for GPS, Proc. 13th Annu. PTI Meet., Washington (1982) pp. 609–630
- 5.53 F. Vannicola, R.L. Beard, J.D. White, K. Senior, M. Largay, J.A. Buisson: GPS block IIF atomic frequency standard analysis, Proc. 42nd Annu. PTI Meet., Reston (2011) pp. 181–196
- 5.54 R.T. Dupuis, T.J. Lynch, J.R. Vaccaro, E.T. Watts: Rubidium frequency standard for the GPS IIF program and modifications for the RAFSMOD program, Proc. ION GNSS 2010, Portland (2010) pp. 781–788
- 5.55 Y.G. Gouzhva, A.G. Gevorkyan, V.V. Korniyenko: Atomic frequency standards for satellite radio navigation systems, Proc. 45th IEEE FCS, Los Angeles (1991) pp. 591–593
- 5.56 F. Droz, P. Rochat, Q. Wang: Performance overview of space rubidium standards, Proc. 24th EFTF, Noordwijk (2010) pp. 1–6
- 5.57 P. Waller, F. Gonzalez, S. Binds, I. Sesia, I. Hidalgo, G. Tobias, P. Tavella: The in-orbit performance of GIOVE clocks, IEEE Trans. Ultrason. Ferroelectr. Freq. Contr. **57**(3), 738–745 (2010)
- 5.58 J.D. White, F. Danzy, S. Falvey, A. Frank, J. Marshall: NTS-2 cesium beam frequency standard for GPS, Proc. 8th Annu. PTI Meet., Washington (1977) pp. 637–664
- 5.59 Symmetricom: *Datasheet 4415 Digital Cesium Frequency Standard* (Symmetricom, San Jose 2003)
- 5.60 S. Fairheller, J. Purvis, R. Clark: The Russian GLONASS system. In: *Understanding GPS – Principles and Applications*, ed. by E.D. Kaplan (Artech House, Boston, London 1996) pp. 439–465
- 5.61 Y.G. Gouzhva, A.G. Gevorkyan, A.B. Bassevich, P.P. Bogdanov, A.Y. Tyulyakov: Comparative analysis of parameters of GLONASS spaceborne frequency standards when used onboard and on service life tests, Proc. 47th IEEE FCS, Salt Lake City (1993) pp. 65–70
- 5.62 A. Bassevich, B. Shebshaevich, A. Tyulyakov, V. Zholnerov: Onboard atomic clocks GLONASS: Current status and future plans, Proc. ION GNSS 2007, Sess. F6a, Fort Worth (2007) pp. 1–11
- 5.63 R.F.C. Vessot, M.W. Levine: A test of the equivalence principle using a space-borne clock, Gen. Relat. Gravit. **10**, 181–204 (1979)
- 5.64 R.L. Easton: The hydrogen maser program for NAVSTAR GPS, Proc. 8th Annu. PTI Meet., Washington (1976) pp. 3–12
- 5.65 J.D. White, A.F. Frank, V.J. Folen: Passive maser development at NRL, Proc. 12th Annu. PTI Meet., Greenbelt (1981) pp. 495–514
- 5.66 H.T.M. Wang: Subcompact hydrogen maser atomic clocks, Proc. IEEE **77**(7), 982–992 (1989)
- 5.67 L. Mattioni, M. Belloni, P. Berthoud, I. Pavlenko, H. Scheda, Q. Wang, P. Rochat, F. Droz, P. Mosset, H. Ruedin: The development of a passive hydrogen maser clock for the galileo navigation system, Proc. 34th Annu. PTI Meet., Reston (2003) pp. 161–170
- 5.68 P. Berthoud, I. Pavlenko, Q. Wang, H. Schweda: The engineering model of the space passive hydrogen maser for the European global navigation satellite system Galileo, Proc. IEEE FCS 17th EFTF 2003, Tampa (2003) pp. 90–94
- 5.69 Q. Wang, P. Mosset, F. Droz, P. Rochat, G. Busca: Verification and optimization of the physics parameters of the onboard Galileo passive hydrogen maser, Proc. 38th Annu. PTI Meet., Reston (2007) pp. 81–94
- 5.70 A. Jornod, D. Goujon, D. Gritti, L.G. Bernier: The 35 kg space active hydrogen maser (SHM-35) for ACES, Proc. IEEE FCS 17th EFTF 2003, Tampa (2003) pp. 82–85
- 5.71 D. Goujon, P. Rochat, P. Mosset, D. Boving, A. Perri, J. Rochat, N. Ramanan, D. Simonet, X. Vernez, S. Froidevaux, G. Perruchoud: Development of the space active hydrogen maser for the ACES mission, Proc. 24th EFTF, Noordwijk (2010) pp. 1–6
- 5.72 J.D. Prestage, S. Chung, T. Le, M. Beach, L. Maleki, R.L. Tjoelker: One-liter Hg ion clock for space and ground applications, Proc. IEEE FCS 17th EFTF 2003, Tampa (2003) pp. 1089–1091
- 5.73 R.L. Tjoelker, J.D. Prestage, L. Maleki: The JPL Hg+ extended linear ion trap frequency standard: Status, stability, and accuracy prospects, Proc. 28th

- 5.74 Annu. PTI Meet., Reston (1997) pp. 245–254
- 5.75 T. Ely, J. Seubert, J. Bell: Advancing navigation timing, and science with the deep space atomic clock, SpaceOps 2014 Conf., Pasadena (AIAA, Reston 2014) pp. 1–19
- 5.76 F.J. Gonzalez Martinez: Performance of New GNSS Satellite Clocks, Ph.D. Thesis (Karlsruher Institut für Technologie, Karlsruhe 2013)
- 5.77 A. Baker: GPS Block IIR time standard assembly architecture, Proc. 22rd Annu. PTI Meet., Vienna (1991) pp. 317–324
- 5.78 H. Rawicz, M. Epstein, J. Rajan: The time keeping system for GPS block IIR, Proc. 24th Annu. PTI Meet., McLean (1993) pp. 5–16
- 5.79 A. Wu: Performance evaluation of the GPS block IIR timekeeping system, Proc. 28th Annu. PTI Meet., Reston, ed. by L. Breakiron (USNO, Washington 1997) pp. 441–453
- 5.80 F.J.M. Carrillo, A.A. Sanchez, L.B. Alonso: Hybrid synthesizers in space: Galileo's CMCU, Proc. 2nd Int. Conf. Recent Adv. Space Technol., Istanbul (2005) pp. 361–368
- 5.81 D. Felbach, D. Heimbuerger, P. Herre, P. Rastetter: Galileo payload 10.23 MHz master clock generation with a clock monitoring and control unit (CMCU), Proc. IEEE FCS 17th EFTF 2003, Tampa (2003) pp. 583–586
- 5.82 D. Felbach, F. Soualle, L. Stopfkuchen, A. Zenzinger: Clock monitoring and control units for navigation satellites, Proc. IEEE FCS 2010, Newport Beach (2010) pp. 474–479
- 5.83 K. Kovach: New user equivalent range error (UERE) budget for the modernized Navstar Global Positioning System (GPS), Proc. ION NTM 2000, Anaheim (ION, Virginia 2000) pp. 550–573
- 5.84 J. Oaks, M. Largay, J. Buisson, W. Reid: Comparative analysis of GPS clock performance using both code phase and carrier derived pseudorange observations, Proc. 36th Annu. PTI Syst. Appl. Meet., Washington (2004) pp. 431–440
- 5.85 J. Ray, K. Senior: Geodetic techniques for time and frequency comparisons using GPS phase and code measurements, Metrologia **42**, 215–232 (2005)
- 5.86 Z. Deng: Reprocessing of GFZ Multi-GNSS product GBM, IGS Workshop 2016, Sydney (IGS, Pasadena 2016)
- 5.87 F. Gonzalez, P. Waller: GNSS clock performance analysis using one-way carrier phase and network methods, Proc. 39th Annu. PTI Meet., Long Beach (ION, Virginia 2007) pp. 403–414
- 5.88 J. Delporte, C. Boulanger, F. Mercier: Simple methods for the estimation of the short-term stability of GNSS on-board clocks, Proc. 42nd Annu. PTI Appl. Plan. Meet., Reston (ION, Virginia 2010) pp. 215–223
- 5.89 O. Montenbruck, P. Steigenberger, E. Schönemann, A. Hauschild, U. Hugentobler, R. Dach, M. Becker: Flight characterization of new generation GNSS satellite clocks, Navigation **59**(4), 291–302 (2012)
- 5.90 A. Hauschild, O. Montenbruck, P. Steigenberger: Short-term analysis of GNSS clocks, GPS Solut. **17**(3), 295–307 (2013)
- 5.91 E. Griggs, E.R. Kursinski, D. Akos: Short-term GNSS satellite clock stability, Radio Sci. **50**(8), 813–826 (2015)
- 5.92 E. Griggs, E.R. Kursinski, D. Akos: An investigation of GNSS atomic clock behavior at short time intervals, GPS Solut. **18**(3), 443–452 (2014)
- 5.93 J.E. Gray, D.W. Allan: A method for estimating the frequency stability of an individual oscillator, Proc 8th Annu. Symp. Freq. Contr., Fort Monmouth (Electronic Industries Association, Washington 1974) pp. 277–287
- 5.94 K. Senior, J. Ray, R.L. Beard: Characterization of periodic variations in the GPS satellite clocks, GPS Solut. **12**(3), 211–225 (2008)
- 5.95 O. Montenbruck, U. Hugentobler, R. Dach, P. Steigenberger, A. Hauschild: Apparent clock variations of the block IIF-1 (SVN-62) GPS satellite, GPS Solut. **16**(3), 303–313 (2012)
- 5.96 O. Montenbruck, P. Steigenberger, U. Hugentobler: Enhanced solar radiation pressure modeling for Galileo satellites, J. Geod. **89**(3), 283–297 (2015)
- 5.97 G. Petit, B. Luzum: *IERS Conventions (2010)*, IERS Technical Note No. 36 (Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt 2010)
- 5.98 J. Kouba: Improved relativistic transformations in GPS, GPS Solut. **8**(3), 170–180 (2004)
- 5.99 J. Kouba: *A Guide to Using International GNSS Service (IGS) Products* (IGS, Pasadena 2015), <http://kb.igs.org/>
- 5.100 Russian Institute of Space Device Engineering: *Global Navigation Satellite System GLONASS – Interface Control Document*, Vol. 5.1 (Russian Institute of Space Device Engineering, Moscow 2008)
- 5.101 Global Positioning Systems Directorate: *Navstar GPS Space Segment/Navigation User Segment Interfaces, Interface Specification*, IS-GPS-200H, 24 Sep. 2013 (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo 2013)
- 5.102 European GNSS (Galileo) Open Service Signal In Space Interface Control Document, OS SIS ICD, Iss. 1.1, Sep. 2010 (EU 2010)
- 5.103 China Satellite Navigation Office: *BeiDou Navigation Satellite System Signal In Space Interface Control Document – Open Service Signal*, v2.0, Dec. 2013 (China Satellite Navigation Office, Beijing 2013)
- 5.104 JAXA: Quasi-Zenith Satellite System Navigation Service Interface Specification for QZSS, IS-QZSS, V1.4, 28 Feb. 2012 (JAXA, Chōfu 2012)
- 5.105 Indian Space Research Organization: *Indian Regional Navigation Satellite System – Signal In Space ICD for Standard Positioning Service*, version 1.0, June 2014 (Indian Space Research Organization, Bangalore, 2014)
- 5.106 H.M. Smith: International time and frequency coordination, Proc. IEEE **60**(5), 479–487 (1972)
- 5.107 H.M. Smith: International coordination and atomic time, Vistas Astron. **28**(1), 123–128 (1985)
- 5.108 T.J. Quinn: The BIPM and the accurate measurement of time, Proc. IEEE **79**(7), 894–905 (1991)
- 5.109 P. Tavella, C. Thomas: Comparative study of time scale algorithms, Metrologia **28**, 57–63 (1991)

- 5.109 C. Audoin, B. Guinot: *The Measurement of Time* (Cambridge Univ. Press, Cambridge 2001)
- 5.110 ITU: Time scales. In: *Handbook Satellite Time and Frequency Transfer and Dissemination* (ITU, Geneva 2010) pp. 78–91
- 5.111 B. Guinot: Some properties of algorithms for atomic time scales, *Metrologia* **24**(4), 195 (1987)
- 5.112 R.A. Nelson, D.D. McCarthy, S. Malys, J. Levine, B. Guinot, H.F. Fliegel, R.L. Beard, T.R. Bartholomew: The leap second: Its history and possible future, *Metrologia* **38**, 509–529 (2001)
- 5.113 ITU: *Standard-Frequency and Time-Signal Emissions, ITU-R Recommendation TF.460-6* (ITU, Geneva 2002)
- 5.114 BIPM: Values of the differences between UTC and its local representations by individual time laboratories (Bureau International des Poids et Mesure, Sèvres 2016) <ftp://ftp2.bipm.org/pub/tai/publication/utclab/>
- 5.115 ITU: *Time-Scale Notation, ITU-R Recommendation TF.536-2* (ITU, Geneva 2003)
- 5.116 C. Han, Z. Cai, Y. Lin, L. Liu, S. Xiao, L. Zhu, X. Wang: Time synchronization and performance of BeiDou satellite clocks in orbit, *Int. J. Navig. Obs.* **371450**, 1–5 (2013)
- 5.117 K.R. Brown: The theory of the GPS composite clock, *Proc. ION GPS 1991, Albuquerque* (1991) pp. 223–241
- 5.118 A.L. Satin, C.T. Leondes: Ensembling clocks of the Global Positioning System (GPS), *IEEE Trans. Aerosp. Electron. Syst.* **26**(1), 84–87 (1990)
- 5.119 K. Senior, P. Koppang, J. Ray: Developing an IGS time scale, *IEEE Trans. Ultrason. Ferroelectr. Freq. Contr.* **50**, 585–593 (2003)
- 5.120 J. Ray, K. Senior: IGS/BIPM pilot project: GPS carrier phase for time/frequency transfer and timescale formation, *Metrologia* **40**, S270–S288 (2003)
- 5.121 K. Senior: Report of the IGS working group on clock products, 19th Meet. Consult. Comm. Time Freq. Sèvres (BIPM, Sèvres 2012) pp. 219–236

Atmospheric

6. Atmospheric Signal Propagation

Thomas Hobiger, Norbert Jakowski

Global navigation satellite system (GNSS) satellites emit signals that propagate as electromagnetic waves through space to the receivers which are located on or near the Earth's surface or on other satellites. Thereby, electromagnetic waves travel through the ionosphere and the neutral atmosphere (troposphere) which causes signals to be delayed, damped, and refracted as the refractivity index of the propagation media is not equal to one. In this chapter, the nature and effects of GNSS signal propagation in both the troposphere and the ionosphere, are examined. After a brief review of the fundamentals of electromagnetic waves their propagation in refractive media, the effects of the neutral atmosphere are discussed. In addition, empirical correction models as well as the state-of-the-art atmosphere delay estimation approaches are presented. Effects related to signal propagation through the ionosphere are dealt in a dedicated section by describing the error contribution of the first up to third-order terms in the refractive index and ray path bending. After discussing diffraction and scattering phenomena due to ionospheric irregularities, mitigation techniques for different types of applications are presented.

6.1	Electromagnetic Wave Propagation	165
6.1.1	Maxwell Equations.....	166
6.1.2	Electromagnetic Wave Propagation in the Troposphere.....	166
6.1.3	Electromagnetic Wave Propagation in the Ionosphere.....	167
6.2	Troposphere.....	168
6.2.1	Characteristics of the Troposphere	168
6.2.2	Tropospheric Refraction	170
6.2.3	Empirical Models of the Troposphere	172
6.2.4	Troposphere Delay Estimation	174
6.3	Ionospheric Effects on GNSS Signal Propagation	177
6.3.1	The Ionosphere.....	177
6.3.2	Refraction of Transionospheric Radio Waves.....	179
6.3.3	Diffraction and Scattering of GNSS Signals	183
6.3.4	Ionospheric Models.....	184
6.3.5	Measurement-Based Ionosphere Correction	189
	References.....	190

6.1 Electromagnetic Wave Propagation

For a long period in the history of natural sciences, light was considered as the only part of the electromagnetic spectrum. Study of light has been conducted in ancient cultures and continued until the sixteenth and seventeenth centuries when the discussion about the nature of light, whether it can be described as a wave or as a particle, arose. It took another century, before electromagnetic waves other than light were studied. In 1800, William Herschel discovered infrared light and soon Johann Ritter noticed an effect which was later described as ultraviolet radiation. In 1845, Faraday discovered the polarization of light in dependence of the strength of a magnetic field and in the 1860s, James

Maxwell developed four partial differential equations for the electromagnetic field. Maxwell soon realized that two of his equations predicted the existence and behavior of waves in the field. Furthermore, he noticed that such waves travel at a speed which was about the known speed of light. Maxwell's theory survived the groundbreaking discoveries of the next two centuries and his equations still represent one of the most elegant ways to express the fundamentals of electricity and magnetism. Based on these mathematical formulations, one can derive wave propagation characteristics that are valid for all wavelengths, which belong to the electromagnetic spectrum.

6.1.1 Maxwell Equations

Maxwell's equations in their classical form (e.g., [6.1]) can be expressed as

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \quad (6.1)$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}, \quad (6.2)$$

$$\nabla \cdot \mathbf{D} = \rho, \quad (6.3)$$

$$\nabla \cdot \mathbf{B} = 0, \quad (6.4)$$

where the quantities \mathbf{E} and \mathbf{H} are the electric and magnetic field vectors, respectively. \mathbf{D} and \mathbf{B} denote the electric and magnetic flux densities, respectively, \mathbf{J} represents the electric current density and ρ is the volume charge density. Thereby, the ∇ operator is defined in a 3-D Cartesian coordinate system \mathbb{R}^3 , which is spanned by the orthogonal unit vectors $\{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$ as

$$\nabla = \mathbf{e}_x \frac{\partial}{\partial x} + \mathbf{e}_y \frac{\partial}{\partial y} + \mathbf{e}_z \frac{\partial}{\partial z}. \quad (6.5)$$

In addition, \cdot and \times denote the inner and outer vector products, respectively. Together with the constitutive relations

$$\mathbf{D} = \varepsilon \mathbf{E}, \quad (6.6)$$

$$\mathbf{B} = \mu \mathbf{H}, \quad (6.7)$$

one can understand how electric and magnetic fields are generated and altered by each other and by charges and currents. The permittivity ε and permeability μ are related to the electric (χ) and magnetic (χ_m) susceptibilities of the material by

$$\varepsilon = \varepsilon_0(1 + \chi), \quad (6.8)$$

$$\mu = \mu_0(1 + \chi_m), \quad (6.9)$$

where ε_0 and μ_0 are the corresponding values in vacuum. Solving the classical Maxwell equations requires sophisticated methods and in many cases results in a numerical solution of the coupled partial derivative equations. However, if certain properties about the medium through which the electromagnetic waves propagate are known, simplification of above equations can be made and propagation characteristics can be derived more easily. This applies in particular for trans-troposphere and trans-ionosphere electromagnetic wave propagation.

6.1.2 Electromagnetic Wave Propagation in the Troposphere

Electromagnetic waves that are transmitted from GNSS satellites pass the ionosphere first, before entering the

neutral atmosphere, in particular, the troposphere. In both media, electromagnetic signals are delayed and refracted, which causes an excess path delay that needs to be corrected in order to realize accurate positioning and timing applications. In order to find a solution for the wave propagation in the troposphere (Sect. 6.2), one can assume an isotropic, nonconducting and neutral medium. Thus, $\mathbf{J} = \mathbf{0}$ and $\rho = 0$ can be applied to (6.2) and (6.3), respectively. As shown by, for example, [6.2], one can derive the wave equations for the electric and magnetic field as

$$\nabla^2 \mathbf{E} = \mu \varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} = \frac{n^2}{c^2} \frac{\partial^2 \mathbf{E}}{\partial t^2}, \quad (6.10)$$

$$\nabla^2 \mathbf{B} = \mu \varepsilon \frac{\partial^2 \mathbf{B}}{\partial t^2} = \frac{n^2}{c^2} \frac{\partial^2 \mathbf{B}}{\partial t^2}, \quad (6.11)$$

where

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} \quad (6.12)$$

is the speed of light in vacuum and

$$n = \sqrt{\frac{\varepsilon \mu}{\varepsilon_0 \mu_0}} \quad (6.13)$$

introduces the concept of the refractivity index. Thus, the knowledge of the refractivity index at any given location along the wave path allows to derive the propagation characteristics, in particular delay and damping of the electromagnetic wave. Before explaining these two phenomena exceeded on the electromagnetic wave, one needs to recall that the refractivity index n is close to one. Thus, many publications introduce the so-called refractivity

$$N = (n - 1) \cdot 10^6, \quad (6.14)$$

which avoids the usage of numbers that distinguish only very little from unity. Since the refractivity index n is a complex number, refractivity N consists of a real and an imaginary part. According to, for example, [6.3], N can be expressed as

$$N = N_0 + N'(f) + jN''(f), \quad (6.15)$$

where the second and third terms reflect the frequency dependence of the refractivity and $j = \sqrt{-1}$ denotes the imaginary unit. Since variations of N over one wavelength are negligible for the troposphere in any of the GNSS bands, propagation effects of real and imaginary parts of N can be dealt with separately. The real part of the refractivity, that is, $(N_0 + N'(f))$

causes electromagnetic waves to be refracted and delayed. Thus, a good model of this propagation behavior, respectively, the possibility to estimate such delays within post-processing will be a crucial issue for high accurate positioning applications. As described in [6.4], one can model the real part of the refractivity by

$$N = \sum_i \left(A_i(f) \rho_i + B_i(f) \frac{\rho_i}{T} \right), \quad (6.16)$$

where ρ_i is the density of the i -th atmospheric gas, T denotes the absolute temperature, and A_i and B_i are constants which are usually determined from experiments. If the composition and distribution of atmospheric gases were known with sufficient accuracy, one might be able to derive an empirical model for N which can be used to deal with troposphere path delays in post-processing of GNSS data. This approach will be picked up and discussed further in Sect. 6.2.2.

The third remaining term in (6.15), that is, $N''(f)$, belongs to the imaginary part of the complex refractivity and is the reason for signal damping, also called attenuation or absorption. In general, $N''(f)$ can be related to the so-called absorption coefficient $\alpha(f)$ by

$$\alpha(f) = 1 \cdot 10^6 \frac{4\pi f N''(f)}{c}. \quad (6.17)$$

Based on this relation, it is possible to compute the receiving power P of a signal which has been sent from transmitter Tr, propagates through the atmosphere along the path S and captured at the receiver Rcv. In doing so, one obtains

$$P = P_0 \exp \left(- \underbrace{\int_{\text{Tr}}^{\text{Rcv}} \alpha(f) ds}_{\kappa(S,f)} \right), \quad (6.18)$$

where P_0 denotes the received signal power in a lossless medium. Although absorption does not directly impact phase and group delay measurements, it has an impact on the quality of the GNSS measurements. Satellite signals tracked at low elevation angles have long propagation paths through the atmosphere, which increases the opacity $\kappa(S,f)$, that is, the integral in (6.18), and thus attenuates the signals stronger than observations taken in the zenith direction.

6.1.3 Electromagnetic Wave Propagation in the Ionosphere

In order to solve the Maxwell equations for electromagnetic waves in the ionosphere, it is necessary to have more information about the physical properties of the medium, especially since \mathbf{J} is not equal to zero. In general, a magnetic plasma, like the ionosphere, is an anisotropic and birefringent medium, which requires the usage of tensor notation for the conductivity $\tilde{\sigma}$ and the dielectric coefficient $\tilde{\epsilon}$. Following the derivation from [6.5] one defines:

- The plasma frequency for electrons

$$f_p^2 = \frac{e^2 n_e}{4\pi^2 m_e \epsilon_0}. \quad (6.19)$$

- The electron gyro (or synchrotron) frequency

$$f_g = \frac{e}{2\pi m_e} \mathbf{B}_{\oplus}. \quad (6.20)$$

- And the collision frequency of the electrons ν .

Where e and m_e are the charge and mass of the electron, respectively, n_e is the free electron density, \mathbf{B}_{\oplus} is Earth's magnetic field vector, and ϵ_0 is the vacuum permeability defined already in the prior section. This leads to the abbreviations

$$X = \frac{f_p^2}{f^2}, \quad Y = \frac{f_g}{f}, \quad Z = \frac{\nu}{f}, \quad (6.21)$$

or

$$\tilde{X} = \frac{X}{1 + jZ}, \quad \tilde{Y} = \frac{Y}{1 + jZ}. \quad (6.22)$$

These yield the conductivity tensor $\tilde{\sigma}$ (in a Cartesian coordinate system when the z -axis is assumed to be in the direction of \mathbf{B}_{\oplus})

$$\tilde{\sigma} = 2\pi f \epsilon_0 \mathbf{j} \begin{pmatrix} \frac{\tilde{X}}{1 - \tilde{Y}^2} & \frac{j\tilde{X}\tilde{Y}}{1 - \tilde{Y}^2} & 0 \\ -\frac{j\tilde{X}\tilde{Y}}{1 - \tilde{Y}^2} & \frac{\tilde{X}}{1 - \tilde{Y}^2} & 0 \\ 0 & 0 & \tilde{X} \end{pmatrix}. \quad (6.23)$$

Moreover, the plasma can be treated neutral if the length of scale is bigger than the Debye length [6.6]. Together with Ohm's law

$$\mathbf{i} = \tilde{\sigma} \mathbf{E}, \quad (6.24)$$

where \mathbf{i} is the current density, one finds the dielectric tensor. This can be expressed as

$$\tilde{\epsilon} = \begin{pmatrix} 1 - \frac{\tilde{X}}{1 - \tilde{Y}^2} & \frac{j\tilde{X}\tilde{Y}}{1 - \tilde{Y}^2} & 0 \\ -\frac{j\tilde{X}\tilde{Y}}{1 - \tilde{Y}^2} & 1 - \frac{\tilde{X}}{1 - \tilde{Y}^2} & 0 \\ 0 & 0 & 1 - \tilde{X} \end{pmatrix}. \quad (6.25)$$

Applying the dielectric tensor to Maxwell's equations and carrying out a few mathematical transformations one finds the expression

$$\mathbf{n} \times (\mathbf{n} \times \mathbf{E}) = -\tilde{\epsilon} \mathbf{E}, \quad (6.26)$$

where the refractivity index vector $\mathbf{n} = (c/f)\mathbf{k}$ has been assigned to the wave vector \mathbf{k} . As discussed in detail in [6.6] the Appleton–Hartree equation

$$n^2 = 1 - \frac{\tilde{X}(1 - \tilde{X})}{1 - \tilde{X} - \frac{\tilde{Y}^2}{2} \pm \sqrt{\frac{\tilde{Y}^4}{4} + \tilde{Y}_L^2(1 - \tilde{X})^2}} \quad (6.27)$$

serves as a solution for (6.26). Two new expressions were introduced in (6.27) by splitting \tilde{Y} into two components. The longitudinal component $\tilde{Y}_L = \tilde{Y} \cos \Theta$ and the transversal component $\tilde{Y}_T = \tilde{Y} \sin \Theta$ account for the angle between the propagation direction and Earth's magnetic field. Again, the refractivity index is a com-

plex quantity and the following properties of a magnetic plasma, such as the ionosphere, can be found:

- *Dispersive*: The index of refraction depends on the used frequency and it can be shown that the group velocity differs from phase velocity.
- *Absorptive*: The index of refraction is a complex number and the imaginary part, called extinction coefficient, describes the energy absorption. This process is dissipative as wave energy is converted into heat through collision processes.
- *Birefringent*: The index of refraction has two distinct values, which suggests the possibility of two ray paths; each one characterized by different phase and group velocities.
- *Anisotropic*: Each of the two indices of refraction is a separate function of the orientation of the normal to the surface of constant wave phase with respect to the background (uniform) magnetic field.

One of the biggest challenges for any system operating in the radio-frequency band is the determination of the propagation velocity of its signal. If such waves propagate through vacuum, the traveled distance would just be the product of speed of light in vacuum with the propagation time between sender and receiver. When signals travel through a magnetic plasma like the ionosphere, phase propagation speed accelerates and group velocity is slowed down.

6.2 Troposphere

Although the terms *atmosphere* and *troposphere* appear to be interchangeably, there exists a clear definition (Sect. 6.2.1) which suggests that the usage of the latter expression is more suitable when dealing with GNSS signal propagation effects. In general, the troposphere is the lowest portion of the Earth's atmosphere where approximately 80% of the atmosphere's mass and 99% of its water vapor and aerosols can be found. Since these constituents and their distribution are of great importance for GNSS signal propagation, in most cases the term *troposphere* is used to describe signal delays, attenuation, and scintillation effects. However, one should keep in mind that about 25% of the *tropospheric delay* is caused by gases which are located above the troposphere, in particular gases in the tropopause and the stratosphere.

6.2.1 Characteristics of the Troposphere

The word troposphere originates from the Greek expression *tropos* which can be translated as *change*. This

reflects the fact that turbulent mixing plays an important role in the troposphere's structure and behavior. Most of the phenomena we associate with day-to-day weather occur in the troposphere. In general, the troposphere starts at the Earth's surface and can reach to a height of 20 km above sea level at maximum. However, in most regions of the world the troposphere reaches only up to roughly 10 km (Fig. 6.1). The prominent feature of this domain is that the temperature generally decreases with altitude at a constant lapse rate of -7 to -5 K/km of altitude. At higher latitudes in winter and at nighttime, a temperature inversion layer in a height between 0.5 and 2 km could exist before the constant lapse rate of the troposphere starts. The boundary between the top of the troposphere and the stratosphere is called the tropopause, which is a region of approximately constant or less varying temperature in a height between 8 and 12 km (Fig. 6.1).

According to the World Meteorological Organization, the tropopause is defined as the lowest level at which the lapse rate decreases to 2 K/km or less, pro-

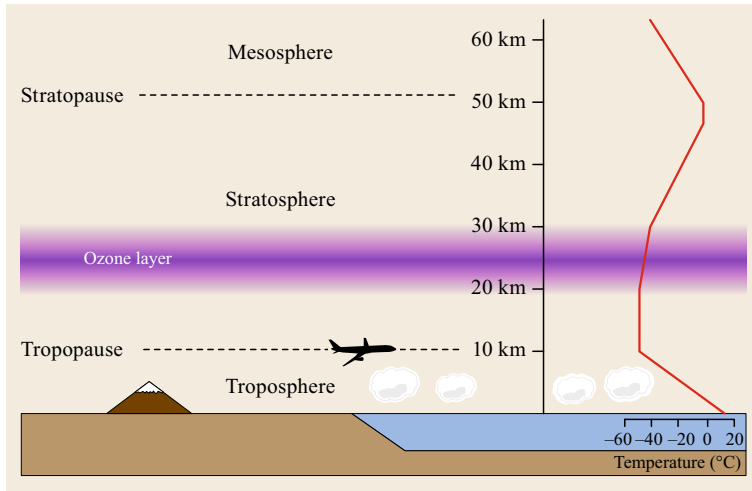


Fig. 6.1 The troposphere is the lowest layer of Earth's atmosphere where weather happens and most of the clouds can be found. The layer up is the stratosphere, followed by the mesosphere. Distinction between the layers is made with vertical temperature lapse rates (see *curve* on the right side) as discussed in Sect. 6.2.1

vided that the average lapse rate between this level and all higher levels within 2 km does not exceed 2 K/km. The height of the tropopause depends on latitude, season, and whether it is day or night. Near the equator, the tropopause is about 20 km above the sea level. In winter near the poles the tropopause is much lower at about 7 km height. In the lower stratosphere, the rate of change of temperature gradually reverses to a positive temperature lapse rate of about 1–2 K/km. Such a slow, nonuniform increase exists through the whole height range of the stratosphere up to a height of roughly 50 km. At this altitude, which is called stratopause, the lapse rate reverses again and temperature will be around 0°C. The region above the stratopause is called mesosphere that is characterized by a negative temperature lapse rate causing the temperature to drop to approximately –90°C at a height of 90 km, where the mesopause marks the end of this atmospheric domain (Fig. 6.1). Knowing that the atmospheric pressure in the mesosphere is relatively small (0.02–1 hPa) and its behavior strongly follows the barometric formula (Sect. 6.2.3), one can understand that the small propagation delay contribution of the mesosphere can be neglected or well modeled if needed.

As shown in the next section, the wet and dry constituents of the atmosphere affect the propagation delay of GNSS signals in different ways. Pressure profiles of water vapor differ much from dry pressure, which follows a more deterministic behavior. Moreover, water vapor is confined to the troposphere, below about 10 km and most of it is found below 4 km. Other than the dry air components, the spatial and temporal distribution of water vapor is highly variable, which makes it difficult to explain phenomena related to this constituent with an

empirical or climatological model. As discussed later in Sect. 6.2.4, GNSS post-processing allows to measure integrated water vapor by estimating its contribution together with the other unknown parameters.

On the other hand, dry atmosphere constituents are only varying little over temporal and spatial scales of hours and kilometers, respectively. Table 6.1 lists the molar weights and percentage on the total volume of the most prominent dry air constituents. Thus, propagation effects exerted from the dry air part can be well modeled and compensated with the help of empirical models. Besides wet and dry air, the atmosphere also contains aerosols (water droplets, ice crystals, salt grains, and dust particles). However, since they do not impact on GNSS signal propagation such constituents are not further discussed here.

Table 6.1 Composition of the *clean dry* air in the standard US model throughout the troposphere. N₂, O₂, and Ar represent 99.96% of the total volume and its composition is quite homogeneous and constant. CO₂ is the only constituent which vary at ground level between day and night by up to a factor of 2

Constituent	Molecular mass (kg/kmol)	Percentage in total volume
N ₂	28.013	78.084
O ₂	32.000	20.946
Ar	39.948	0.934
CO ₂	44.010	0.033
Ne	20.183	0.0018
He	4.003	0.0005
Kr	83.8	0.0001
CH ₄	16.043	0.0002
H ₂	2.016	0.00005
N ₂ O	44.013	0.00005

6.2.2 Tropospheric Refraction

As discussed in the prior section, the abundance of dry atmosphere constituents is remarkably homogeneous and constant. Moreover, the *wet part* of the atmosphere, that is, water vapor, is the only constituent which has a significant dipole moment that can influence the propagation of electromagnetic waves. Since contributions from liquid water droplets do not exceed millimeter order, they can be ignored for GNSS measurements and one can re-write (6.16) as the sum of three main constituents, that is,

$$N(f) = k_1(f) \frac{p_d}{T} Z_d^{-1} + k_2(f) \frac{p_w}{T} Z_w^{-1} + k_3(f) \frac{p_w}{T^2} Z_w^{-1} \quad (6.28)$$

as described in [6.7]. Thereby, the pressure of dry air p_d controls the first contribution and wet air pressure p_w governs the second and third contribution. In doing so, the compressibility factors for dry Z_d and wet Z_w air describe how these constituents differ from an ideal gas. Although (6.28) is still frequency dependent, one should keep in mind that by neglecting liquid water vapor, one has implicitly restricted the applicability to the electromagnetic propagation in the lower microwave spectrum. The compressibility factor Z_i , is the ratio of the molar volume $V_{m,i}$ of a gas i to the molar volume of an ideal gas $\hat{V}_{m,i}$ at the same temperature T and pressure p . This relation can be expressed as

$$Z_i = \frac{V_{m,i}}{\hat{V}_{m,i}} = \frac{pV_{m,i}}{RT} = \frac{pM_i}{\rho_i RT}, \quad (6.29)$$

where R is the universal gas constant, M_i is the molar mass of the i -th constituent, and ρ_i the corresponding density. The usage of empirically determined values makes it possible to derive a model for the incompressibilities, which depend only on temperature and the pressure. In doing so, [6.8] obtains

$$Z_d^{-1} = 1 + p_d \left[57.97 \cdot 10^{-8} \left(1 + \frac{0.52}{T} \right) - 9.4611 \cdot 10^{-4} \frac{T_c}{T^2} \right] \quad (6.30)$$

and

$$Z_w^{-1} = 1 + 1650 \frac{p_w}{T^3} (1 - 0.0131 T_c + 1.75 \cdot 10^{-4} T_c^2 + 1.44 \cdot 10^{-6} T_c^3), \quad (6.31)$$

where T_c is the temperature in degree Celsius, whereas T is the absolute temperature in Kelvin. If the coefficients $k_i(f)$ are known with sufficient accuracy, one can

derive wet and dry refractivity models which only depend on temperature, pressure and water vapor content.

In the following, only electromagnetic propagation effects in a frequency range below 40 GHz will be considered. Since all microwave-based space-geodetic techniques, including GNSS, operate in this frequency domain, a set of three coefficients, k_1 , k_2 , and k_3 , without explicit frequency dependence will be sufficient to link between temperature, pressure, and water vapor pressure and total refractivity. Thus, (6.28) simplifies to

$$N = k_1 \frac{p_d}{T} Z_d^{-1} + k_2 \frac{p_w}{T} Z_w^{-1} + k_3 \frac{p_w}{T^2} Z_w^{-1}. \quad (6.32)$$

Various measurements in laboratories and other studies have been carried out to determine accurate coefficients for the k_i values. A thorough discussion of the history of these measurements and the best average values according to [6.9] can be found in [6.10]. Table 6.2 lists these values, which should be used for GNSS processing and for conversion between nonhydrostatic delay and integrated water vapor.

The first term in (6.32) is sometimes referred as dry refractivity, whereas the second and third are called wet refractivity. However, in most of the literature another distinction is made, which expresses (6.32) under consideration of the relation (6.29) as

$$N = \underbrace{k_1 \frac{R}{M_d} \rho}_{N_h} + \underbrace{k_2' \frac{p_w}{T} Z_w^{-1} + k_3 \frac{p_w}{T^2} Z_w^{-1}}_{N_w}, \quad (6.33)$$

where

$$k_2' = k_2 - k_1 \frac{M_w}{M_d}. \quad (6.34)$$

Here, N_h is called hydrostatic refractivity and the sum of the latter two terms in (6.33) is called nonhydrostatic or wet refractivity. The introduction of mean molar masses of dry air, M_d , and wet air, M_w , enable the usage of the factor k_2' for a better representation of (6.33). This way of expressing the refractivity budget has the advantage that the hydrostatic refractivity depends only on the total density of the air, which can be easily deduced from ground-based pressure measurements (Sect. 6.2.3). On the other side, the nonhydrostatic (or wet) part can be

Table 6.2 Suggested values of the refractivity coefficients and their uncertainty (after [6.9])

Coeff.	Value	Unit
k_1	77.6890 ± 0.015	K/hPa
k_2	71.2952 ± 10	K/hPa
k_3	$375\,463 \pm 3000$	K ² /hPa

related to temperature and water vapor pressure, a concept which allows the determination of highly variable water vapor variations by means of GNSS as described in Chap. 38.

Although a tiny frequency dependence of the refractivity exists in theory, simulations can show that such an influence on troposphere delays is below 0.2 mm for all elevation angles. Thus, the troposphere can be treated as a nondispersive medium, that is, $dN/df = 0$. This feature simplifies the derivation of easy-to-apply troposphere propagation models, troposphere delay effects are identical for group- and phase delay observations. However, it also implies that troposphere excess delays cannot be removed by dual- or multifrequency measurements like in the case of trans-ionospheric propagation (Sect. 6.3.5). Therefore, it is necessary to estimate troposphere delays by a sophisticated functional model within the parameter adjustment process. In doing so, it is important to understand how to relate between refractivity N , respectively, index of refractivity n and the troposphere signal delay. According to Fermat's principle, the path taken by an electromagnetic wave minimizes the total delay between the transmitter T and the receiver R. In doing so, the ratio between propagation time and speed of light (in vacuum) is called electric path length L and can be expressed as

$$L = \int_R^T n(s) ds, \quad (6.35)$$

where ds is an infinitesimal distance along the true ray path. In case of vacuum propagation, the index of refractivity equals one and the straight line which connects T and R becomes the ray-path. This straight line length \overline{TR} is called geometric distance G . As for the troposphere, $n > 1$ causes electromagnetic waves to propagate slower than in vacuum and continuous refraction causes the ray-path to be bent in according to Fermat's principle. Thus, one can determine the troposphere delay ΔL as

$$\Delta L = \int_R^T n(s) ds - G. \quad (6.36)$$

If we recall the definition of refractivity (6.14), this equation can be reformulated as

$$\Delta L = 10^{-6} \int_R^T N(s) ds + S - G, \quad (6.37)$$

where $S = \int_R^T ds$ is the geometric length of the true (bended) propagation path. Together with (6.33) it is

now possible to write

$$\Delta L = 10^{-6} \left(\int_R^T N_h(s) ds + \int_R^T N_w(s) ds \right) + \underbrace{S - G}_{\Delta_g}, \quad (6.38)$$

which makes it clear that the total troposphere delay can be interpreted as the sum of three contributions. The first part is called hydrostatic delay

$$\Delta L_h = 10^{-6} \int_R^T N_h(s) ds \quad (6.39)$$

and the second contribution is referred to as wet delay

$$\Delta L_w = 10^{-6} \int_R^T N_w(s) ds. \quad (6.40)$$

Other than ΔL_h and ΔL_w , which consider that electromagnetic waves are slowed down in the atmosphere, the last contribution Δ_g originates from the bending effect, which causes signals to travel a longer path than the straight connection between T and R. For practical reasons when dealing with mapping functions (Sect. 6.2.3), Δ_g is included in the hydrostatic part.

Since GNSS observations are taken at any arbitrary azimuth and elevation angle, one needs to know the hydrostatic and wet delay contributions at each of these observing directions in order to be able to remove the troposphere excess delay which biases the observations. However, as shown in Sect. 6.2.4, one is able to estimate troposphere parameters when assuming a relation between zenith troposphere delays and those observed at arbitrary elevations and the azimuth directions. Therefore, it is useful to introduce the zenith hydrostatic delay (ZHD)

$$\text{ZHD} = 10^{-6} \int_{h_0}^{h_\infty} N_h(z) dz \quad (6.41)$$

and the zenith wet delay (ZWD)

$$\text{ZWD} = 10^{-6} \int_{h_0}^{h_\infty} N_w(z) dz. \quad (6.42)$$

In doing so, vertical integration needs to be carried out from height h_0 to the upper height of the atmosphere h_∞ . As shown in the next section, empirical zenith delay models can be rather easily established due to the strong vertical alignment of the atmosphere.

6.2.3 Empirical Models of the Troposphere

Recalling (6.33), it is obvious that the hydrostatic delay depends only on the total density of the air. Following the idea of a hydrostatic equilibrium, a motionless fluid has zero net force on it and thus the sum of the forces in a given direction must be opposed by an equal sum of forces in the opposite direction. This hydrostatic balance can be expressed in a 1-D (vertical) case as

$$\frac{\partial p}{\partial z} + \rho(z)g(z) = 0, \quad (6.43)$$

where $g(z)$ is the total gravity acceleration at a given height z . Integrating vertically provides the pressure p_0 at a given (geopotential) height h_0 by

$$p_0 = \int_{h_0}^{\infty} \rho(z)g(z)dz = g_{\text{eff}} \int_{h_0}^{\infty} \rho(z)dz, \quad (6.44)$$

where the effective gravity g_{eff}

$$g_{\text{eff}} = \frac{\int_{h_0}^{\infty} \rho(z)g(z)dz}{\int_{h_0}^{\infty} \rho(z)dz} \quad (6.45)$$

was introduced. In doing so, g_{eff} is the gravity acceleration, which is representative for an atmosphere with density variation $\rho(z)$. As gravity is monotonically decreasing with height, one can interpret g_{eff} as the gravity acceleration at the centroid height h_c of the atmospheric column, that is,

$$h_c = \frac{\int_{h_0}^{\infty} \rho(z)zdz}{\int_{h_0}^{\infty} \rho(z)dz}. \quad (6.46)$$

According to [6.11], the approximation

$$h_c = (0.9 h_0 + 7300 \text{ m}) \quad (6.47)$$

holds for all latitudes and seasons with an uncertainty of about ± 400 m. Thus, one can compute h_c and use this value to derive the effective gravity acceleration g_{eff} for any given location.

Pressure and Temperature Information

Almost all empirical models that are able to model troposphere delays for the analysis of GNSS data require, the user to input pressure values at the locations where the observations have been taken. Thus, unless a pressure sensor is installed at the site the user is left with the problem to obtain accurate pressure information, which can be used for the computation of hydrostatic and wet delay models. Berg [6.12] derived the simple empirical model

$$p = 1013.25 \text{ hPa} (1 - 2.25 \cdot 10^{-5} \text{ m}^{-1} h)^{5.225}, \quad (6.48)$$

which provides pressure values for a given orthometric [6.13] height h . Hopfield [6.14] presented another, more sophisticated model for computing empirically pressure at a given height h . His approach considers a temperature lapse rate $\alpha = 4.5 \text{ K/m}$ and reads as

$$p = 1013.25 \text{ hPa} \left(\frac{T_k - \alpha h}{T_k} \right)^{\frac{g}{R_d \alpha}}, \quad (6.49)$$

where $T_k = 293.16 \text{ K}$, that is, 20°C , was assumed to be the temperature at the sea level. Together with the value for the mean gravity acceleration $g = 9.7867 \text{ ms}^{-2}$ and the dry air gas constant $R_d = 0.287 \text{ kJ/K/kg}$ users are able to derive pressure in the case that no in-situ measurements were available at the ground site. However, with millimeter accurate GNSS observations, such models turned out to be insufficient for post-processing. Thus, users either need to extract accurate pressure values from numerical weather models (NWMs) or rely on more sophisticated empirical models for the computation of such values. For the latter choice, a variety of models have been presented recently, which are capable to provide accurate pressure information that does not bias or degrade the GNSS solutions. UNB3m [6.15], global pressure and temperature (GPT [6.16]), and GPT2 [6.17] are the most prominent models. The latter one is also recommended by the current conventions [6.18] of the International Earth Rotation and Reference Systems Service (IERS).

UNB3m was the first empirical model that could be used to obtain a good guess of meteorologic parameters in the case that no meteo sensor is available at a GNSS site. This model not only provides pressure and temperature and relative humidity, but also outputs other meteorologic information that might be useful for GNSS processing. However, as UNB3m was mainly derived from US standard atmosphere [6.19] data the model did not represent temporal variations other than those modeled by an annual signal. This drawback was partially overcome by the introduction of the GPT, which represents the meteorologic quantities in the form of spherical harmonic coefficients of degree and order 9. Although temporal variations are still limited to an annual term, the major advantage of GPT is the data source on which the coefficients were fitted. Other than UNB3m, GPT uses numerical weather model data to determine the empirical model coefficients. Thus, GPT had been suggested in the IERS conventions [6.18] for processing of space geodetic data before it was replaced by its successor, that is, GPT2. Like GPT, GPT2 is derived from numerical weather model data. However, spherical harmonics are replaced with gridded coefficients, semiannual terms are included and a more sophisticated up- and down-continuation algo-

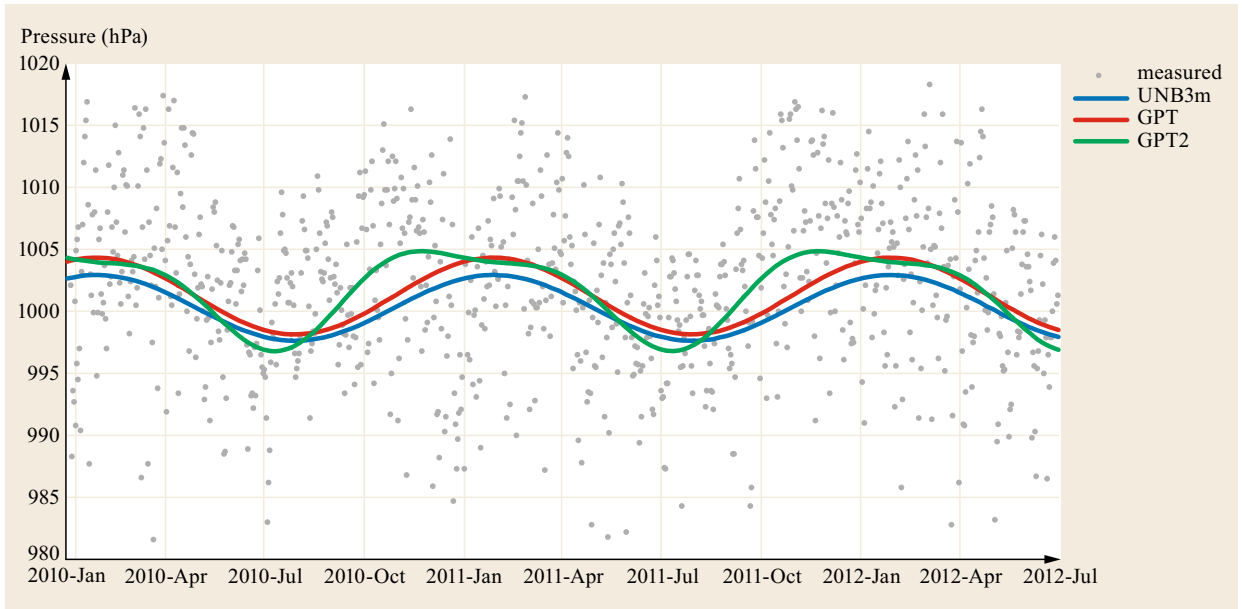


Fig. 6.2 Daily pressure values (dots) at 12 h local time (i. e., 3 h UT) at Koganei, Tokyo, Japan, are plotted together with the pressure predictions from the UNB3m (blue), GPT (red), and GPT2 (green) models

rhythm is being used to obtain temperature and pressure at the user's station. A comparison of modeled pressure values with predictions from three different pressure models is shown in Fig. 6.2.

The Saastamoinen Hydrostatic Delay Model

Since hydrostatic delays are solely dependent on pressure, one can derive an accurate model for this part of the atmosphere delay quite easily. Considering (6.44) and replacing the vertical integration over the total air density in (6.33) yield

$$\text{ZHD} = 10^{-6} k_1 \frac{R p_0}{M_d g_{\text{eff}}}, \quad (6.50)$$

which allows to relate the hydrostatic zenith delay directly to ground pressure p_0 . In doing so, one needs to have an accurate empirical model for g_{eff} as defined in (6.45). References [6.11, 20] describe how to find such a relation based on a standard gravity model. They suggest using

$$g_{\text{eff}} = 9.7840 \text{ m/s}^2 \kappa(\varphi, h_0) \quad (6.51)$$

with

$$\begin{aligned} \kappa(\varphi, h_0) = & 1 - 0.00266 \cos(2\varphi) \\ & - 0.28 \cdot 10^{-6} \text{ m}^{-1} h_0, \end{aligned} \quad (6.52)$$

where φ is the latitude of the site and h_0 has to be the geopotential height of the receiver (6.44).

Wet Delay Models

Other than the hydrostatic delay contribution, wet delays are less accurately predictable from ground-based sensor data. Due to the high spatial and temporal variability of water vapor, one can only derive a model with limited capability to predict the ZWD

$$\begin{aligned} \text{ZWD} = & 10^{-6} \left(\int_{z_0}^{z_\infty} \left(k'_2 \frac{p_w}{T} Z_w^{-1} \right) dz \right. \\ & \left. + \int_{z_0}^{z_\infty} \left(k_3 \frac{p_w}{T^2} Z_w^{-1} \right) dz \right). \end{aligned} \quad (6.53)$$

at a given site. According to [6.10], the first term is about 60 times smaller than the second one in (6.53). Various models for ZWD have been proposed for space geodetic applications. For example, *Saastamoinen* [6.11] focused on the ideal gas law and developed the simple formula

$$\text{ZWD} = 0.0022768 (1255 + 0.05 T_s) \frac{p_{ws}}{T_s}, \quad (6.54)$$

which provides a reasonable estimate of the wet delay contribution when temperature T_s and water vapor pressure p_{ws} are measured at the site. *Hopfield* [6.14] followed another approach, leading to the basic relation

$$\text{ZWD} = \frac{10^{-6}}{5} N_w(h_s) h_w, \quad (6.55)$$

where $N_w(h_s)h_w$ is the wet air refractivity at the site (at height h_s) and $h_w = 11\,000$ m is the assumed height of the tropopause under which water vapor can be found. Besides various other models, the suggestion from [6.21] is also worth to be mentioned here as it suggests the simple linear relation

$$\text{ZWD} \approx 0.217 \frac{p_w}{T}, \quad (6.56)$$

which can be used as a good approximation for a-priori ZWDs. As discussed in [6.10], one can derive from this model an approximation based on ground meteo data, that is,

$$\text{ZWD} \approx 748 \frac{p_{ws}}{T_s^2}. \quad (6.57)$$

For applications with even less accuracy requirements on ZWD, one can use the rule of thumb estimate

$$\text{ZWD} \approx \frac{p_{ws}}{100} \quad (6.58)$$

when the surface water vapor pressure p_{ws} is given in units of hPa. All empirical models for wet delay prediction (6.91)–(6.96) have in common that they rely solely on surface or site dependent data. Such models will give reasonable estimates of ZWD for standard meteorologic conditions, but fail to predict ZWDs with centimeter accuracy for weather situations that deviate from such an assumption. Given the rather low accuracy of such models, ZWDs are usually estimated whenever enough satellites are tracked in order to separate this delay contribution from station coordinates and clock parameters. For applications that target lower accuracy, the wet delay contribution might be either neglected or approximated by one of the models listed above.

6.2.4 Troposphere Delay Estimation

Empirical troposphere models which were discussed in the prior section have been only defined for the zenith direction. Thus, one needs to have a good mathematical model which relates between zenith troposphere delays and the actual delay at a given elevation angle. In addition, if the accuracy requirements of a GNSS application do not permit to apply empirical troposphere delay corrections, one needs to estimate such excess delay together with the other unknown parameters. Both issues can be dealt with when introducing the concept of mapping function, which are discussed here.

Separation of Parameters

GNSS observations, that is, code and carrier phase measurements, are affected by troposphere delays in

the same way. Other than ionosphere delays, which can be canceled out (at first-order) by dual-frequency measurements (Sect. 6.3.5), propagation through the neutral atmosphere (troposphere) does not cause dispersive delays on GNSS measurements. Although various approaches for modeling troposphere delays have been discussed in the prior sections, it is important to recall that all these models are available in the zenith direction. Thus, if troposphere excess delays should be corrected by such models only, one needs to have a mathematical approach that relates between a delay in the zenith direction ($\tau(E = 90^\circ)$) and the one observed at a given elevation angle ($\tau(E)$). Defining a so-called mapping function (Sect. 6.2.4)

$$M(E) \approx \frac{\tau(E)}{\tau(E = 90^\circ)}, \quad (6.59)$$

which approximates the fraction on the right with sufficient accuracy, allows to apply a-priori zenith troposphere corrections for GNSS observations at arbitrary elevation angles. However, since such models, in particular those for the wet troposphere delays, are limited in their accuracy, one needs to choose another approach to remove troposphere delays when realizing highly precise applications by means of GNSS. Thus, instead of only correcting troposphere delays by a-priori models, one can estimate troposphere excess delays by making use of the mapping function approach. Assuming that mapping functions for hydrostatic ($M_h(E)$) and wet ($M_w(E)$) are provided in a convenient form, one can apply the following two-step processing strategy:

1. Since hydrostatic zenith troposphere delays can be computed with high accuracy, one can use, for example, the Saastamoinen model (Sect. 6.2.3) for deriving ZHD at a given GNSS site. Together with a proper mapping function ($M_h(E)$), this information can be transformed into slant troposphere delays at any elevation angle. Such delays can be applied as corrections in post-processing, respectively, in the parameter adjustment process. Since hydrostatic delays are the dominating contributor to the troposphere delay budget, more than 85% of any total troposphere delay can be accounted for by this approach.
2. The remaining troposphere delays, which are thought to be attributable to wet delays need to be estimated within the parameter adjustment process. This is usually realized by adding a time-dependent representation of ZWD as additional parameters which are estimated together with the other unknown parameters. In order to separate wet troposphere delay from the other parameters, one takes

advantage from the mapping function approach which introduces an elevation dependent scaling factor, that is, $M_w(E)$, for each observation.

The mapping function $M(E)$ follows $\approx 1/\sin(E)$ in first-order and thus provides a unique partial derivative for least squares extended Kalman filter or other adjustment methods. Figure 6.3 discusses parameters, that is, station position, clock, and troposphere, which can be separated together due to different elevation dependent behavior. Thus, when estimating troposphere parameters together with the other unknowns, one should make sure that a sufficient number of observations at different elevation angles are taken in order to properly separate the different physical signals.

Mapping Functions

The simplest form of any mapping function can be obtained through the rather crude assumption of a plane earth together with a horizontally stratified atmosphere where no variations of the refractivity occur within a layer. Ignoring bending effects, the path delay will then simply be proportional to the propagation path length in each layer and the mapping function will be identical to the cosecant function, that is,

$$M(E) = \csc E = \frac{1}{\sin E}. \quad (6.60)$$

This simple approximation works surprisingly well, having an error of about 1% (corresponding to approximately 2 cm for the zenith component of total delay at a sea level site) for an elevation angle of 20° . However, in order to improve the positioning results and take into account observations from lower elevations, such a simple mapping function approximation does not

work well for applications with a higher demand on accuracy. Obviously, the cosecant mapping function can be improved by introducing a spherical earth instead of a plane model. In order to have a slightly better treatment of the hydrostatic and wet components separately, *Chao* [6.22] derived the following expressions that also depend on the elevation angle only

$$M_h(E) = \frac{1}{\sin E + \left(\frac{0.00143}{\tan E + 0.0445} \right)} \quad (6.61)$$

and

$$M_w(E) = \frac{1}{\sin E + \left(\frac{0.00035}{\tan E + 0.017} \right)}. \quad (6.62)$$

They were recommended to be used down to elevation angles of approximately 15° . There are also other models suggested by *Chao* [6.22] with similar accuracy [6.23, 24]. Further improvements can be made by using models for the pressure, temperature, and humidity profiles. Several different versions of such mapping functions were developed and published in [6.14, 25–28]. The advantage with this type of mapping function is that the result is a closed-form solution where the input parameters are surface meteorology and parameters describing the characteristics of the pressure, temperature, and humidity profiles.

In general, it is important to mention that hydrostatic mapping function does not only model the ratio between slant and ZHDs, but also account for the extra delay caused by signal bending in the atmosphere. This convention ensures that wet mapping functions, deliver only the wet component of the troposphere delay, when being applied in parameter adjustment.

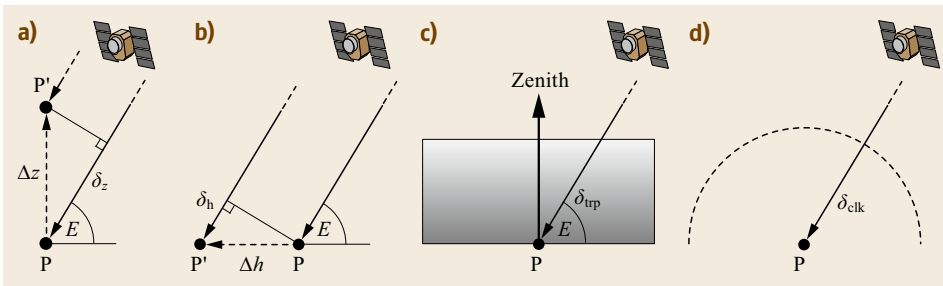


Fig. 6.3a–d Elevation dependency of different target parameters. **(a)** A change in the vertical Δz leads to an extra delay path which can be expressed as $\delta_z(E) = \Delta z \sin(E)$. **(b)** A change in the horizontal component Δh causes a delay of $\delta_h(E) = \Delta h \cos(E)$ (Note: Here a two-dimensional (2-D)-case, with the horizontal displacement aligned toward the satellite, is being depicted for better readability. For a three-dimensional (3-D) case one needs to add the azimuth dependency for the two components.) **(c)** Troposphere delays depend on the mapping function, which in first-order can be approximated by $\delta_{trp}(E) \sim \delta_{trp}(E = 90^\circ) 1/\sin(E)$. **(d)** A clock offset appears to be isotropic, that is, independent from the elevation angle

Mapping Functions of the Continued Fraction Form

The use of the continued fraction form for the mapping function was suggested by *Marini* [6.29] and reads as

$$M(E) = \frac{1}{\sin E + \frac{a}{\sin E + \frac{b}{\sin E + \frac{c}{\sin E + \dots}}}} \quad (6.63)$$

Based on ray-tracing through standard atmospheres models various authors determined the coefficients a , b , and c as functions of pressure, water vapor pressure, and temperature on the Earth surface, or from temperature gradients and the height of the troposphere. The MTT (Mapping Temperature Test) mapping function [6.30] can be seen as an extension of this mathematical approach, but is modified so that the value of the mapping function is one in the zenith direction. It is formulated as

$$M(E) = \frac{1 + \left[\frac{a}{(1+b/(1+c))} \right]}{\sin E + \left[\frac{a}{\sin E + (b/(\sin E + c))} \right]} \quad (6.64)$$

The parameters a , b , and c depend here on ground surface temperature, the station latitude, and the orthometric height of the station, respectively. We note that the only parameter that needs to be measured is the temperature at the ground. The values of the empirical parameters are also derived here through ray tracing, not using a standard atmospheres model but a set of radiosonde data. A potential problem (or difficulty) with the Harvard–Smithsonian Center of Astrophysics (CfA-2.2, [6.20]), the Ifadis, and the MTT mapping functions is that they all use ground surface observations of meteorological parameters. Temperature inversions often introduce errors in these mapping functions where a too low surface temperature implies an underestimation of the temperature at higher altitudes. We also note that in addition to the work involved to obtain surface observations there is always the risk that poor instrumentation can introduce gross measurement errors. These potential measurement errors are often ignored in the analysis carried out in order to compare the accuracy of different mapping functions. This was the main motivation for *Niell* to formulate the *Niell* mapping function (NMF) [6.31]. The advantage of the NMF is that its parameters do not need any observations of ground meteorology. It is sufficient to know the time of the year, the latitude, and the height of the site. One may suspect that the NMF could be less accurate at extreme sites since it cannot be tuned to local weather conditions by means of surface observations. We have, however, not seen any such findings reported up until now. The

NMF was, for example, tested on data from Chajnantor, a very dry site at a height of 5 km in Chile, with good results [6.32].

Mapping Functions Based on NWP Model Results

The ultimate mapping function is of course only possible to obtain for a complete knowledge of the 3-D refractivity field in the atmosphere above the actual GNSS site. A method using data from either a climatological model or numerical weather prediction (NWP) analyses was developed and presented by *Rocken* et al. [6.33]. They refer to this method as *direct* mapping since its value at any specified elevation angle is obtained through a ray-tracing analysis based on vertical profiles of pressure, temperature, and humidity. In the study presented, one mapping function per day was computed and compared to ray-trace results from collocated radiosonde launches. Of course, this method can be developed further using the full 3-D-fields in order to also take possible azimuthal asymmetries into account. It will, however, mean that the number of ray-traces has to be significantly larger. *Niell* [6.34] developed the first mapping functions based on (6-hourly) data from NWM. An azimuthal symmetric mapping function for the hydrostatic delay was developed following the three term continued fraction form but with the additional parameter being the 200 hPa geopotential height obtained from NWM analysis. Since the hydrostatic mapping function is based on the height of the 200 hPa pressure level, it is called the isobaric mapping function (IMF). The IMF offered an improvement of approximately a factor of two compared to the NMF [6.34, 35]. The Vienna mapping functions (VMFs) [6.36] and VMF1 [6.37] have been developed along the same lines as the IMF but focusing on direct ray-tracing of the NWP reanalyses from the European Centre for Medium Range Weather Forecasting (ECMWF). The gridded and site-wise mapping function coefficients can be downloaded at [6.38]. The result from ray-tracing at an elevation angle of 3° is used to determine the value of the parameter a in the NMF type of formula. They recommend the use of direct mapping functions rather than using the intermediate parameters, such as the geopotential height at the 200 hPa pressure level. The University of New Brunswick (UNB) [6.39] is also calculating VMF1 coefficients, using data from the NWM of the National Centers for Environmental Prediction (NCEP) and an NVM of the Canadian Meteorological Centre (CMC). They call it UNB-VMF1 and the parameters can be retrieved from [6.40]. Also the Deutsche GeoForschungsZentrum Potsdam (GFZ) is working on a realization of the VMF1 and other developments related to ray-tracing and mapping functions.

The global mapping functions (GMFs) [6.41] serve as a blind, or climatologic, counterpart to the VMF1. Recently, there was an update to GMF, called GPT2 [6.17], also providing improved blind pressure values at the sites. GMF and GPT2 are in the tradition of the mapping function [6.31].

In recent years, there has been considerable research on mapping functions, for example, the *adaptive mapping function* described in [6.42]. At some point, a clear separation between mapping functions and ray-tracing is no longer possible with the new developments. For any observation, the slant and the corresponding zenith delay can be determined by ray-tracing, the ratio being then called slant factor or mapping functions. The use of ray-tracing for tropospheric delay correction has, for example, been studied in [6.43–45].

Gradients

Mapping functions, as discussed in the prior section, work under the assumption of an azimuthal symmetric atmosphere around the GNSS site. However, due to local and regional climatic and weather conditions, atmosphere delays at a constant elevation angle will slightly vary with the azimuth direction. In order to account for such an effect, the so-called gradients need to be estimated together with the wet troposphere delays for utmost high precise positioning applications. Currently,

two gradient mapping functions have been suggested to model the dependence on the azimuth angle A . *MacMillan* [6.46] proposes to use the simple form

$$M_{\text{gr}}(A, E) = M_h(E) \cot E (G_N \cos A + G_E \sin A) \quad (6.65)$$

for estimating the two horizontal gradients, that is, G_N in the north–south direction and G_E in the east–west direction. One has to use the hydrostatic mapping function $M_h(E)$ as described in [6.35]. In [6.47], the gradient mapping function

$$M_{\text{gr}}(A, E) = \frac{1}{\sin E \tan E + C} (G_N \cos A + G_E \sin A) \quad (6.66)$$

is suggested, which does not depend on another (hydrostatic) mapping function, but considers elevation dependency by its own model. *Chen and Herring* [6.47] suggest to use $C = 0.0031$ and $C = 0.0007$, if one estimates hydrostatic and wet gradients separately. However, such a separation is hardly used nowadays, and only total gradients are estimated within post-processing. In doing so, one should use $C = 0.0032$ as discussed in [6.30].

6.3 Ionospheric Effects on GNSS Signal Propagation

To understand transionospheric radio wave propagation and related effects on GNSS signals, we will briefly describe the main features of the ionosphere which are relevant for the subsequent discussion.

6.3.1 The Ionosphere

The ionosphere is the ionized part of the Earth's atmosphere, ranging from about 50 km up to about 1000 km height. Above the so-called transition region at around 1000 km height the ionized and co-rotating atmosphere is usually called plasmasphere or protonosphere reaching up to the plasmopause height at about 3–5 Earth radii in the equatorial plane. The term *plasmopause* characterizes the boundary to the outer magnetosphere. The ionospheric plasma is basically generated by solar radiation, which dissociates and ionizes neutral molecules and atoms. Since the energy of solar radiation in the visible (V) and infrared (IR) wavelength ranges is too low to ionize the neutral gas, they can reach the Earth's surface (Fig. 6.4).

The main constituents (i.e., O, O₂, N, N₂, NO) of the upper atmosphere are ionized by the solar radiation

in the far and extreme ultraviolet (EUV) wavelength ranges ($\lambda < 130$ nm) as well as by high energetic solar x-rays. To a less extent, cosmic rays and energetic particles originating from the solar wind may also contribute to the ionospheric ionization. Consequently, the plasma is composed by a variety of different atomic and molecular ions interacting in a complex way by photochemical reactions. Basic processes can be described by the continuity and energy equations and equations of motion for the individual charged particles taking into account that the total number of ions is equal to the number of electrons in the ionospheric plasma indicating quasi-neutrality of the ionosphere. The fundamental continuity equation for the electrons is given by

$$\frac{\partial n_e}{\partial t} = Q_e - L_e - \nabla \cdot (n_e v_e), \quad (6.67)$$

where n_e is the electron density, t is the time, Q_e is the rate of electron production, L_e is the rate of electron loss, and v_e is the mean velocity of the electrons [6.6]. The divergence term stands for the net loss/gain of plasma due to transport processes. It is evident that density, composition, and temperature of the neutral gas

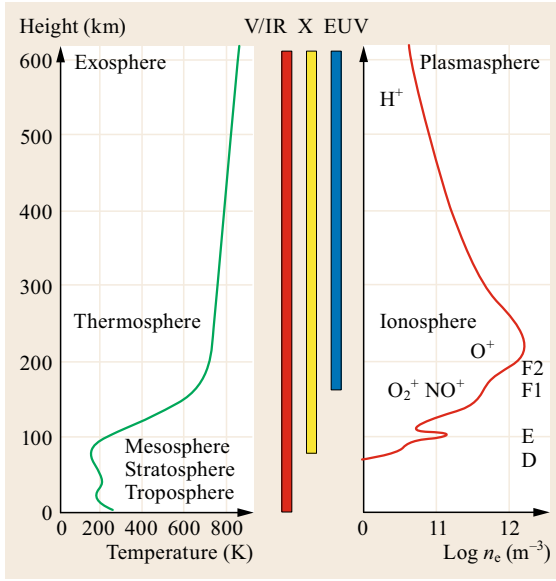


Fig. 6.4 Vertical structure of the electron density of the ionosphere (right) in comparison with the neutral atmosphere temperature (left) and solar radiation penetration depths (middle)

have a severe impact on production, loss, and transport terms in the continuity equation.

The strong coupling of the ionosphere with the thermosphere is due to the fact that the plasma is a trace gas in the neutral gas with an ionization degree of about 10^{-3} at the F2 layer height. Thus, it becomes clear that thermospheric density affects all terms in the continuity equation (6.67). To give only a few examples: the higher the atomic oxygen density – the higher the photo-production; the higher the molecular densities – the higher the loss term. Transport processes due to diffusion and neutral winds depend also on thermospheric conditions. Thus, collisions between neutrals and charged particles as ions and electrons impact neutral winds by ion-drag. On the other hand, neutral winds may lift up the ionospheric plasma up and down along the geomagnetic field lines. Separation of electrons and ions by neutral winds may cause electric fields that are able to drive electric currents. Electric conductivity of magnetospheric origin may generate strong currents in the lower ionosphere where the electric conductivity maximizes around 100 km height (Fig. 6.4). Electric currents, on the other hand, may significantly heat the thermosphere, thus changing composition and temperature and finally all the constituents of the continuity equation. Composition changes, modifying the ratio between molecular and atomic constituents, have an essential impact on the ionization during ionospheric storms.

The complex dynamics of production, loss, and motion of the ionospheric plasma including strong coupling in particular with the thermosphere and magnetosphere lead to a typical vertical structure of the ionospheric electron density as shown in the right panel of Fig. 6.4.

The different ionospheric layers (named as D, E, F1, and F2 in order of increasing altitude) characterize regions where specific processes dominate. The E layer was named by E.V. Appleton (the physics Nobel prize winner in 1947) and designates the region where electrical conductivity and electric currents peak. The ionization at altitudes around 100 km is strongly impacted by high energetic particles and radiation such as x-rays accompanying solar flare eruptions which are associated with significant impact on terrestrial radio wave propagation.

The vertical electron density distribution can basically be described by Chapman's theory [6.48]. Considering a horizontally stratified isothermal layer of a one-component gas, which is ionized by a monochromatic beam of solar radiation at an incidence angle χ , the height dependence of the electron density n_e is given by the Chapman layer function

$$n_e = N_0 \exp \left(\frac{1}{2} [1 - z - \sec \chi \exp(-z)] \right) \quad (6.68)$$

with

$$z = \frac{h - h_0}{H}. \quad (6.69)$$

Here, N_0 denotes the peak electron density of the layer, h is the height above Earth's surface, h_0 the peak electron density height, and H the pressure scale height of the neutral gas in the background. Although the aforementioned assumptions simplify the physics in (6.68), the Chapman layer formula describes the basic features of the vertical structure of the ionospheric electron density very well.

As has already been shown in Sect. 6.1, the electron density is the most important parameter characterizing the refractive index for radio waves in the ionosphere. This is also valid for the integral of the vertical electron density, which is called as total electron content (TEC) and counts all electrons in a columnar cylinder of unit area. The vertical total electron content (VTEC) is defined by

$$\text{VTEC} = \int n_e dh, \quad (6.70)$$

whereas the often used slant total electron content (STEC) measured along a slant ray path s is accordingly

defined by

$$\text{STEC} = \int n_e ds. \quad (6.71)$$

TEC is commonly measured in TEC units (1 TECU = 10^{16} electrons/m²).

Since the ionizing solar radiation varies up to more than 50% within a typical 11 years solar cycle in particular at shorter wavelengths, the total ionospheric ionization, that is, VTEC, is strongly related to the solar cycle as illustrated in Fig. 6.5. The solar radio flux measured at 10.7 cm wavelength in Ottawa is a well-established index ($F_{10.7}$) of the solar activity due to its high correlation with the ionizing solar radiation in the EUV range.

Due to the strong coupling with the thermosphere and magnetosphere, which are also impacted by electromagnetic radiation and particles originating from the solar wind, the ionospheric behavior is very dynamic and closely related to space weather conditions as discussed in Chap. 39. Consequently, the impact on GNSS signal propagation depends significantly on space weather conditions. On the other hand, dual frequency GNSS measurements enable an effective monitoring of TEC and the ionospheric electron density distribution for studying solar terrestrial relationships.

6.3.2 Refraction of Transionospheric Radio Waves

To reach customers at the ground, signals of global navigation satellite systems (GNSS) must travel through the ionosphere. It is evident that the electromagnetic field of radio waves interacts with charged particles whose motion is controlled by the geomagnetic field. The degree of interaction is described by the refractive index n which has been derived in the late 1920s

and early 1930s by Appleton, Lassen, and Hartree by applying Maxwell's theory to the ionospheric plasma (Sect. 6.1). For further reading see [6.5, 49] and references therein.

When taking into account the real ionosphere, that is, considering the fact that due to the presence of the geomagnetic field the ionosphere is an anisotropic medium and collisions between neutrals and charged particles (ions and electrons) are allowed, the refractive index n is given by the Appleton–Hartree formula [6.5, 49]. Since GNSS frequencies (L-band) are well above the collision frequencies, which are in the order of a few kHz, the complex formula can be simplified by neglecting collision terms. The refractive index n is then given by the equation

$$n^2 = 1 - \frac{2X(1-X)}{2(1-X) - Y^2 \sin^2 \Theta \pm [Y^4 \sin^4 \Theta + 4(1-X)^2 Y^2 \cos^2 \Theta]^{\frac{1}{2}}}, \quad (6.72)$$

where

$$X = \frac{f_p^2}{f^2}, \quad (6.73)$$

$$Y = \frac{f_g}{f}, \quad (6.74)$$

and Θ is the angle between the ray path and the geomagnetic field induction B (Fig. 6.6), f is the radio wave frequency, f_p is the plasma frequency, and f_g is the gyrofrequency of electrons that gyrate clockwise around the field lines in the field direction.

The plasma frequency f_p is the resonance frequency of an electron gas of density n_e excited by a radio wave

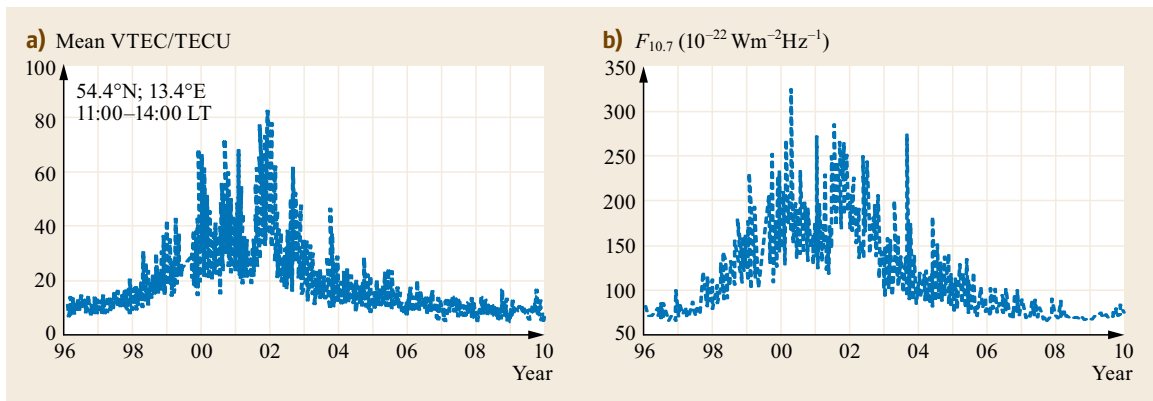


Fig. 6.5a,b GNSS-based TEC measurements obtained at 54.4°N and 13.4°E between 11:00 and 14:00 LT (a) in comparison with the solar radio flux index $F_{10.7}$ (b)

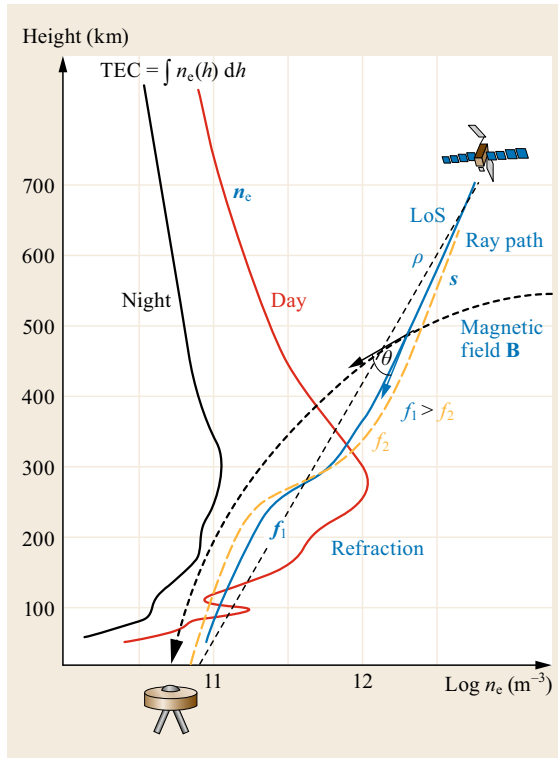


Fig. 6.6 Scheme of transionospheric radio wave propagation at two frequencies f_1 and f_2 in the presence of the geomagnetic field B

and is given by

$$f_p^2 = \frac{n_e e^2}{4\pi^2 \epsilon_0 m_e} \quad (6.75)$$

where e is the electron charge, m_e is the electron mass, and ϵ_0 is the dielectric constant of vacuum.

The gyrofrequency f_g denotes the frequency at which electrons gyrate around magnetic field lines in the ionosphere (clockwise in the field direction). The gyrofrequency is as a function of the geomagnetic field induction B according to

$$f_g = \frac{eB}{2\pi m_e} \quad (6.76)$$

The gyrofrequency is in the order of ≤ 1.4 MHz near the Earth surface and decreases with radial distance r from the center of the Earth by $1/r^3$. According to (6.72), the ionosphere is a dispersive and anisotropic propagation medium. Since the collision terms have been neglected, the absorption capability of the ionosphere is ignored in this high frequency approach. Absorption phenomena play a significant role in terrestrial radio communication; here they are completely ignored.

Due to the small size of the X and Y terms ($X \approx 10^{-5}$ and $Y \approx 10^{-3}$), the refractive index n can be expanded in inverse powers of frequency. The expansion up to the fourth inverse powers of frequency gives

$$n = 1 - \frac{1}{2}X \pm \frac{1}{2}XY \cos \Theta - \frac{1}{4}X \left[\frac{1}{2}X + Y^2(1 + \cos^2 \Theta) \right] \quad (6.77)$$

In terms of plasma, gyro, and wave frequencies, this equation can be written as

$$n = 1 - \left(\frac{f_p^2}{2f^2} \right) \pm \left(\frac{f_p^2 f_g}{2f^3} \cos \Theta \right) - \left(\frac{f_p^2}{8f^4} \left[\frac{f_p^2}{2} + f_g^2(1 + \cos^2 \Theta) \right] \right) \quad (6.78)$$

where the terms in brackets denote the first-, second- and third-order refraction effects.

The double signs in (6.72), (6.77), and (6.78) indicate double refraction of a radio wave travelling through the nonisotropic plasma of the ionosphere. The anisotropy is due to the presence of the geomagnetic field which leads to a double refraction depending on the direction and strength of the geomagnetic field along the ray path. The upper (+) sign represents the ordinary wave (left-hand side circularly polarized) whereas the negative sign refers to the extraordinary wave (right-hand side circularly polarized) [6.50–52]. Equations (6.77) and (6.78) indicate that $n < 1$, that is, the phase velocity $v = c/n$ is greater than the speed of light in vacuum causing a phase advance.

If the wavelength $\lambda = c/f$ of a radio wave traveling through the ionosphere is much smaller than characteristic spatial scales of the ionosphere S_1 ($\lambda \ll S_1$), principles of geometrical optics can be applied.

This means that the propagation follows Fermat's law of fastest arrival, meaning that the phase integral or eikonal $L = \int n ds$ becomes a minimum [6.5].

Considering the ray path s and the line-of-sight (LOS) ρ as shown in Fig. 6.6, the optical path length can be rewritten as

$$s = \underbrace{\int ds_0}_{\rho} + \underbrace{\int (n-1) ds}_{\Delta s_\varphi} + \underbrace{\int ds - \int ds_0}_{\Delta s_B} \quad (6.79)$$

Here, ρ is the true range between the transmitting satellite and the ground receiver along the line-of-sight (Fig. 6.6), Δs_φ stands for the range error terms measured by phase changes and Δs_B is the optical ray path excess due to bending [6.53].

Whereas the true range ρ shall be determined in positioning, ionosphere probing techniques utilize in particular the residual terms of phase measurements in Δs_φ , which contain the electron density along the ray path. Instead of ionospheric range errors Δs_φ and Δs_B , the ionospheric time delay is defined as

$$t_{DI} = \frac{(\Delta s_\varphi + \Delta s_B)}{c}, \quad (6.80)$$

where c is the velocity of light in vacuum.

When measuring the travel time of the code phase in GNSS praxis, we have to take into account the group refractive index of the radio signals [6.5, 6]. The group refractive index is defined by

$$n_{gr} = n + f \left(\frac{dn}{df} \right). \quad (6.81)$$

Inserting (6.78) for n , we get

$$n_{gr} = 1 + \left(\frac{f_p^2}{2f^2} \right) \mp \left(\frac{f_p^2 f_g}{f^3} \cos \Theta \right) + \left(\frac{3f_p^2}{4f^4} \left[\frac{f_p^2}{2} + f_g^2 (1 + \cos^2 \Theta) \right] \right), \quad (6.82)$$

where the terms in brackets denote the first-, second-, and third-order group refraction effects.

To simplify the subsequent discussion of higher order effects, we define the ionospheric range errors for the carrier phase measurements d_I and code measurements d_{Igr} as follows

$$d_I = \int (1 - n) ds, \quad (6.83)$$

$$d_{Igr} = \int (n - 1) ds. \quad (6.84)$$

These terms represent the ionospheric range errors that have to be used in GNSS observation equations for the code d_{Igr} and carrier d_I phase measurements as discussed in Chap. 19. GNSS applications commonly use the first-order approach defined by (6.76) and (6.81) or by (6.77) and (6.82) according to

$$d_I^{(1)} = \int \frac{f_p^2}{2f^2} ds = \frac{K}{f^2} \int n_e ds, \quad (6.85)$$

with

$$K = \frac{e^2}{8\pi^2 \epsilon_0 m_e} \approx 40.309 \text{ m}^3 \text{ s}^{-2}. \quad (6.86)$$

Neglecting higher order terms in (6.78) and (6.82), the first-order ionospheric effect can effectively be eliminated by a linear combination of dual-frequency measurements (Chap. 20). For precise GNSS measurements that require accuracy at centimeter level and below, the consideration of higher order effects and bending is needed.

Since global positioning system (GPS) signals are right hand polarized [6.54], we consider only the extraordinary wave (corresponding to the lower sign in (6.78) and (6.82)) in the subsequent discussion. Following [6.53], we get

$$d_I = d_I^{(1)} + d_I^{(2)} + d_I^{(3)} = \frac{p}{f^2} + \frac{q}{2f^3} + \frac{u}{3f^4}, \quad (6.87)$$

$$d_{Igr} = d_{Igr}^{(1)} + d_{Igr}^{(2)} + d_{Igr}^{(3)} = \frac{p}{f^2} + \frac{q}{f^3} + \frac{u}{f^4}, \quad (6.88)$$

where the parameters p , q , and u are defined in SI units as

$$p = K \int n_e ds = 40.309 \int n_e ds, \quad (6.89)$$

$$q = 2.2566 \cdot 10^{12} \int n_e B \cos \Theta ds, \quad (6.90)$$

and

$$u = 2437 \int n_e^2 ds + 4.74 \cdot 10^{22} \int n_e B^2 (1 + \cos^2 \Theta) ds. \quad (6.91)$$

Figure 6.7 gives an impression of the maximum size of the different error types. Considering the L1 navigation signal frequency at low elevation at extremely high ionization level (here 250 TECU), first-order ionospheric range errors can exceed 100 m. Assuming that the gyrofrequency f_g defined by (6.76) is usually less than 1.4 MHz ($B = 50 \mu\text{T}$) and the propagation is along the field line ($\Theta = 0$, maximum range error condition) the second-order error should fulfill the condition

$$d_I^{(2)} \leq \frac{5.6 \cdot 10^7}{f^3} \text{ STEC}, \quad (6.92)$$

i. e., $d_I^{(2)}$ should generally be less than 12 cm for L1 (25 cm for L2 and 29 cm for L5) frequency even at low elevation and high solar activity conditions.

The asymmetry can be observed at a selected GPS receiver site in particular in the North–South direction

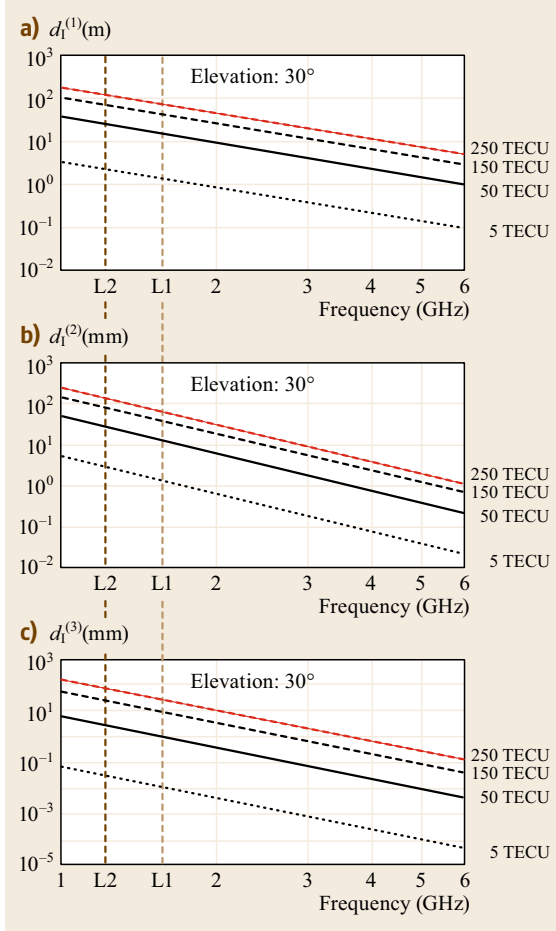


Fig. 6.7a–c Frequency dependence of (a) first-, (b) second- (longitudinal propagation, that is, $\Theta = 0$) and (c) third-order ionospheric range errors for radio waves between 1 and 6 GHz at different levels of vertical TEC for 30° elevation angle of the radio link. GPS frequencies L1 and L2 are marked by dashed lines

as shown e.g., by [6.52, 53, 55, 56]. Due to its systematic character, the effect is meaningful in dual frequency precise geodetic measurements and satellite orbit determination where millimeter accuracy is required [6.56].

The third-order effect defined in (6.93) is more difficult to estimate due to its dependence from the vertical electron density profile shape [6.50, 57, 58]. Following [6.53] one gets

$$d_1^{(3)} \leq \frac{812.3}{f^4} \int n_e^2 ds \approx \frac{534.2}{f^4} N_m F_2 \text{ STEC}, \quad (6.93)$$

where $N_m F_2$ means the peak electron density of the electron density profile (Fig. 6.4). Considering extreme values for $N_m F_2$ and STEC, it can be stated that the

third-order error for L1 frequency is generally less than 6 mm for L1 (16 mm for L2 and 19 mm for L5) frequency, that is, in most cases it can be ignored. It should be mentioned that according to (6.87) and (6.88), the corresponding group delays have to be multiplied by a factor of 2 for the second and by a factor of 3 for the third-order terms. So the group delays at L1 signal should generally be less than 24 cm and 18 mm for the second- and third-order effects, respectively. The corresponding worst case values for the new L5 signal are 58 cm and 57 mm for the second- and third-order effects, respectively.

To complete the discussion of higher order effects, we will now consider the bending effect Δs_B as already sketched in Fig. 6.6 and introduced in (6.79). The bending effect has been estimated so far by analytical approaches [6.58] or by numerical ray-tracing computations (e.g., [6.53, 59–61]).

Based on numerical ray-tracing computations, the approximate relation

$$\Delta s_B = \frac{b_1}{f^4} \left(\frac{1}{\sqrt{1 - b_2 \cos^2 E}} - 1 \right) \text{ STEC}^2 \quad (6.94)$$

(hereafter denoted as JPM94 model) has been established in [6.59] for the excess path length due to bending of GNSS signals. Here $b_1 = 2.495 \cdot 10^8$, $b_2 = 0.859$ and E is the elevation angle in radians. The excess path length Δs_B in (6.94) will be given in millimeters when STEC is measured in TECU and the frequency f is measured in MHz. The model depends only on STEC and therefore can easily be applied.

A more refined approximation, which includes electron density profile parameters such as the atmospheric scale height H and the peak density height $h_m F_2$ is given by

$$\Delta s_B = \frac{7.5 \exp(-2.13E) \text{ STEC}^2}{10^5 f^4 H (h_m F_2)^{1/8}} \quad (6.95)$$

(H&J08 model [6.53]), where Δs_B is measured in meter, STEC is measured in TECU, the frequency f is measured in GHz, scale height H and peak density height $h_m F_2$ are measured in kilometer and elevation E is measured in radians. It has been found that this approach showed the best performance in comparison with ray-tracing computations (Fig. 6.8).

As shown in Fig. 6.9, bending errors can practically be ignored at elevation angles larger than 60° but can exceed the 1 cm level at elevation angles less than 30°. Even under high solar activity conditions, the bending error should generally be less than 5 cm at the L1 frequency (< 14 cm at L2, < 17 cm at L5).

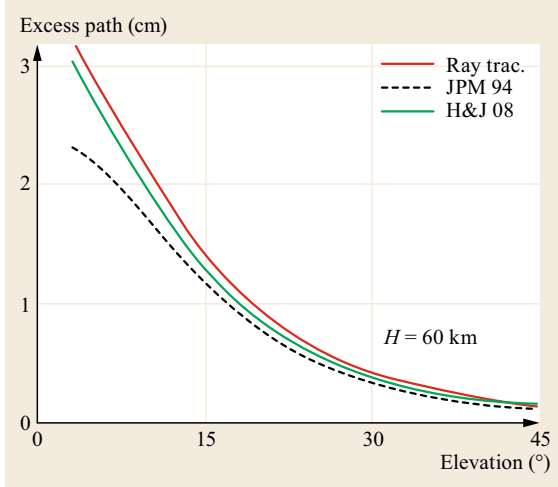


Fig. 6.8 Comparison of JPM94 and H&J08 approximations for the excess path at L2 frequency with ray-tracing computations using a VTEC value of 123 TECU and electron density profile shape parameters $N_m F_2 = 4.96 \cdot 10^{12} \text{ m}^{-3}$, $H = 60 \text{ km}$, and $h_m F_2 = 350 \text{ km}$

6.3.3 Diffraction and Scattering of GNSS Signals

If typical scales of ionospheric electron density variations are comparable with the wavelength, principles of geometric optics are no longer valid. The propagation of radio waves through the ionosphere must then be described by diffraction and scattering theories [6.62, 63]. Here, we confine our attention to describe the phenomenological impact of small scale ionospheric irregularities on radio signals. Such irregularities cause rapid changes in the amplitude and/or phase of radio signals commonly known as radio scintillations [6.64]. Scintillations superposing the signal reduce the accuracy and reliability of radio systems and may even result in a complete loss of lock of the signal. Scintillation effects on satellite signals cover a broad frequency range from 30 MHz up to 10 GHz. Strong scintillations can typically last up to several hours.

To estimate the spatial size of ionospheric irregularities causing scintillations at radio signals of wavelength λ , the first Fresnel zone can be considered. The corresponding radius F_1 is defined by

$$F_1 = \sqrt{\frac{\lambda d_1 d_2}{d_1 + d_2}}. \quad (6.96)$$

Here, d_1 and d_2 denote the distance from the transmitter and receiver, respectively (Fig. 6.10). In the case of GNSS applications, the first Fresnel radius is in the order of about 300 m. Irregularities of this size or smaller

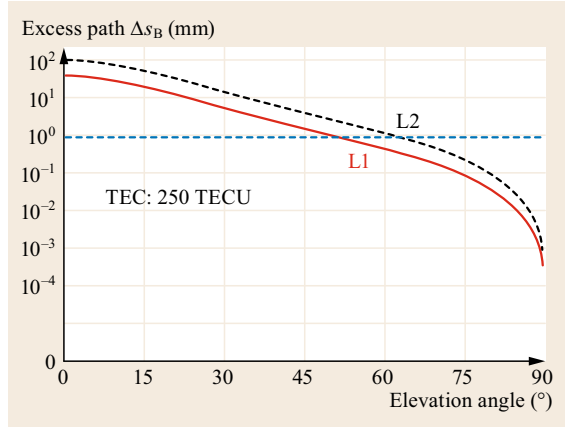


Fig. 6.9 Elevation angle dependence of the excess path length at GNSS frequencies for VTEC = 250 TECU. The 1 mm level is marked by a dashed line

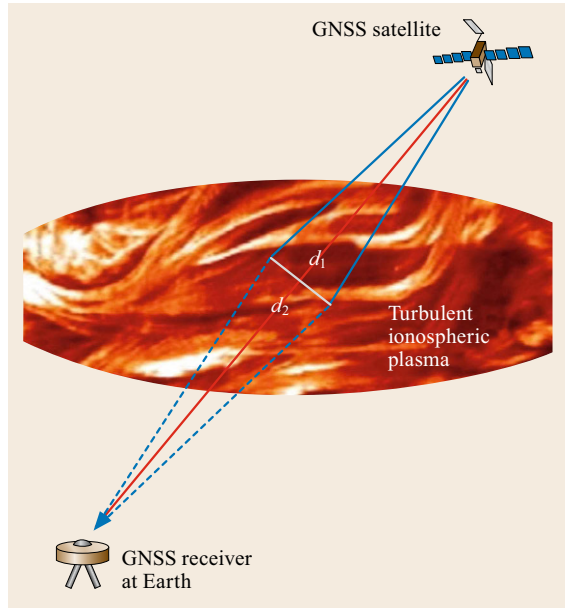


Fig. 6.10 Superposition of diffracted and scattered radio signals at the receiver antenna due to plasma density turbulences

are most effective in producing multiple diffracted and scattered radio waves that interfere at the receiver antenna.

Small-scale electron density irregularities (illustrated in Fig. 6.10) split the primary ray into many different rays causing strong and rapid signal fluctuations at the receiver level. If the fading depth is strong enough, the receiver loses signal tracking, that is, the availability of signals for positioning and navigation might heavily be reduced.

Ionospheric irregularities are closely related to plasma instabilities. One of them is the Rayleigh–Taylor instability (RTI) [6.65]. The RTI describes the behavior of two fluids or plasma volumes moving in the opposite directions. This regularly happens in the low latitude ionosphere in particular during the sunset hours when plasma diffusion is directed downward due to plasma cooling and on the other hand, plasma is uplifted due to an eastward directed electric field. Although the actual geophysical conditions modify the establishment of the RTI, enhanced scintillation activity between sunset and midnight is a well-known phenomenon at low latitudes. Enhanced scintillation activity can also be observed along strong ionization gradients probably initiated by the gradient drift instability [6.66]. Furthermore, irregular precipitation of energetic particles originating from the solar wind may also create chaotic plasma structures resulting at high latitudes in radio scintillations [6.67].

At low latitudes, the so-called equatorial plasma bubbles (EPBs) may be formed by nonlinear plasma processes probably closely related to the RTI. Inside a plasma bubble, the electron density is extremely low (less than 10% of the background density). Hence, there is a sharp gradient of electron density when crossing the surface of an EPB. In TEC data, this is indicated by a rapid drop-off when the ray path enters the EPB and recovery when the ray path leaves the EPB. Whereas EPBs are shaped along magnetic field lines up to more than 1000 km, they are rather thin perpendicular to field lines (up to about 100 km). As RTI is establishing near sunset, EPBs occur and drift eastward at a velocity of about 100–200 m/s [6.68]. The occurrence probability depends on solar activity and on season with highest values around equinoxes over Africa and around solstices at the American sector [6.69].

Low-level geomagnetic activity is anticorrelated with the generation of EPBs, whereas severe magnetic storms may cause enhanced generation of EPBs.

The scintillation strength of received signals is commonly described by the scintillation index S_4 . Other parameters useful for characterizing scintillations are the phase standard deviation σ_φ , the probability and duration of fades, and their depth in the signal strength.

The S_4 index is commonly defined via the signal intensity SI by

$$S_4 = \left(\frac{\langle SI^2 \rangle - \langle SI \rangle^2}{\langle SI \rangle^2} \right)^{1/2}, \quad (6.97)$$

where $\langle \rangle$ means the average value over a 1 min interval. S_4 index values are usually between 0 and 1. Values lower than 0.2 represent low, values around 0.5 medium

and values greater than 0.7 severe scintillation activity. The phase scintillation index widely used is defined by

$$\sigma_\varphi = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\varphi_i - \langle \varphi \rangle)^2}, \quad (6.98)$$

where φ means the signal phase and N the number of observations. Parameters S_4 and σ_φ are commonly defined over a period of 1 min.

The scintillation activity depends on radio frequency. In a first approximation, the scintillation level varies with the inverse of the frequency in the range $1.7 \text{ GHz} < f < 4 \text{ GHz}$.

Severe scintillation conditions may cause loss of lock of the signal thus reducing the availability of signals, that is, the position dilution of precision (PDOP) quality. Summarizing, most severe scintillation effects are observed at and near the equatorial regions and at high latitudes. In the auroral and polar cap latitudes, any significant magnetic storm activity can produce scintillation effects. Usually high-latitude scintillations are not as severe as those measured in the near-equatorial belt. However, they can last for many hours, even days, and are not limited to the local late evening hours as the near-equatorial scintillation effects. The maximum fading depth observed on GPS L1 C/A code signals from the north polar cap region was approximately about 10 dB, whereas in the equatorial anomaly region the fading depth is observed to be as much as 25 dB [6.70]. Typical aspects of scintillation occurrence probability at a selected site are described in scintillation models such as the WideBandModel (WBMOD) [6.71] and the Global Ionospheric Scintillation Model (GISM) [6.72, 73]. The occurrence of strong scintillations is closely related to the solar activity. During the years of maximum solar activity strong scintillation effects on GPS are observed in the equatorial and low-latitude region. Further details of scintillation characteristics including seasonal dependency will be discussed in Chap. 39. It has been shown that high rate GNSS measurements are suitable to monitor scintillations for studying and modeling ionospheric scintillations.

6.3.4 Ionospheric Models

As pointed out in Sect. 6.3.2 and expressed in (6.85), the first-order ionospheric propagation error depends only on the electron density distribution along the ray path. Therefore, ionospheric models that describe the 3-D electron density distribution around the globe as a function of time enable estimating link-related ionospheric propagation errors.

Since the correction term depends only on the integral of the electron density along the ray path (STEC), the availability of a much simpler 2-D model of the vertical TEC (VTEC) can be sufficient for numerous applications. As described in the previous section, the vertical TEC must be transformed to the required slant ray path to finally get STEC.

Generally speaking, when using ionospheric models for correcting ionospheric propagation errors for single frequency measurements, the goodness of correction depends on the quality of the model used. When using a TEC model, the quality of correction depends additionally on the correctness of the applied mapping function. Whereas GPS users rely on a simple TEC model (GPS or Klobuchar model) that is broadcasted to single-frequency users, the European satellite navigation system Galileo offers the internal use of a 3-D model (NeQuick) for single frequency corrections. Both models are described in the following in comparison with two other models currently available for first-order ionospheric corrections. Correction approaches for higher order and bending errors have been discussed already in Sect. 6.3.2.

Mapping Function

Ionospheric correction models provide normalized information in terms of vertical delay or VTEC. If vertical TEC is provided as reference, any link related slant TEC (STEC) can be computed by using a so-called obliquity factor or mapping function $M(E)$ that depends only on the ray path elevation E (Fig. 6.11). Since such

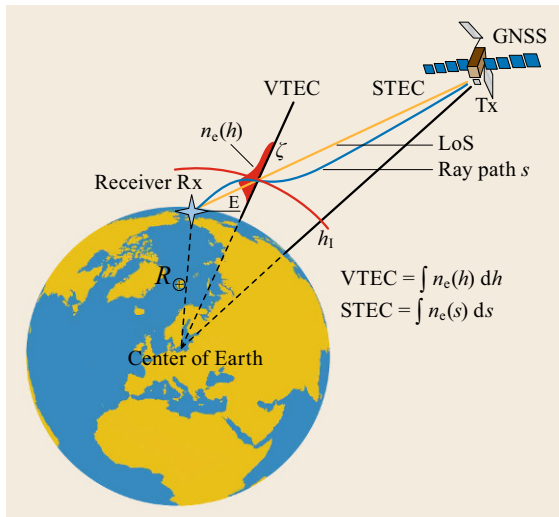


Fig. 6.11 Thin-shell mapping function approach for ionospheric corrections of slant GNSS measurements deduced from vertical TEC (VTEC) and vice versa for mapping VTEC deduced from STEC measurements

a transformation requires additional knowledge about the spatial structure of the ionosphere that is not commonly available, mapping errors result.

The conversion of vertical to STEC information and vice versa is illustrated in Fig. 6.11, where the ionosphere is reduced to a thin shell. Within this simplification, the piercing point of the ray path s from the satellite transmitter Tx to the receiver Rx along the line-of-sight (LOS) with the shell at height h_1 defines geographic coordinates (often called subionospheric point), for which the TEC transformation is valid. The thin-shell mapping function is utilized in most GNSS single-frequency applications, including, for example, satellite-based augmentation systems (SBASs) for aviation.

Assuming a thin-shell-ionosphere and applying simple geometric relationships (Fig. 6.11), the mapping function $M(E)$ for converting VTEC to the corresponding STEC at the piercing point of the ray path s with the ionospheric shell at the height h_1 is given by

$$M(E) = \frac{\text{STEC}}{\text{VTEC}} = \frac{1}{\cos \zeta} = \left[1 - \left(\frac{R_{\oplus} \cos E}{R_{\oplus} + h_1} \right)^2 \right]^{-1/2}, \quad (6.99)$$

where R_{\oplus} is the Earth radius, h_1 is the height of the thin shell representing the ionosphere and E is the elevation angle. The ionospheric shell height is usually fixed within the height interval of 350–450 km. In SBAS like the Wide Area Augmentation System (WAAS) and the European Geostationary Navigation Overlay Service (EGNOS) the thin-shell height is fixed at 350 km [6.74]. After fixing the ionospheric shell height h_1 at a certain value, the thin-shell mapping function is solely dependent on the ray path elevation angle. Thus it ignores the spatial structure of the ionosphere, in particular also horizontal gradients.

According to [6.74], a thin-shell mapping function may introduce vertical range errors of up to 10 m, which increase by a factor of 2–3 at low elevation. Such large mapping errors are due to strong deviations from ionospheric equilibrium conditions in conjunction with strong horizontal gradients of ionospheric ionization as occurred during the Halloween storm by the end of October 2003. Since slant range errors of 15 m or more violate the protection level of SBAS, the navigation service becomes unavailable under such conditions.

Several attempts have been made to improve the TEC mapping in single frequency applications in both the directions, that is, for correcting slant GNSS measurements by given VTEC and for converting measured STEC into VTEC in monitoring systems. It is obvi-

ous from Fig. 6.11 that the mapping improves when the shell height h_1 used in (6.99) is adapted to height variations of the center of ionospheric ionization, that is, when h_1 follows the variations of the peak density height $h_m F_2$ as shown by Sakai et al. [6.75]. Unfortunately, such specific information is not available on a regular base. To improve the TEC mapping from STEC measurements, one option is to separate the ionosphere into different spherical layers with specific mapping functions of type (6.99) [6.75, 76]. Another option is to apply tomographic methods that provide 3-D estimations of the ionospheric electron density [6.77]. The tomographic method is only applicable if sufficient data coverage is available. The above mentioned multilayer shell model as well as the tomographic mapping method have been developed specifically for generating vertical TEC maps as precise as possible. They are not applicable to correct slant GNSS measurements from vertical delay or VTEC without additional information. To improve the correction of slant GNSS measurements by using VTEC without additional information, a new mapping function has been proposed in [6.78], which takes benefit from current knowledge of the typical vertical structure of the ionosphere as described by the Chapman layer in (6.68). The method is applied along the ray path trace through numerous ionospheric shells of incremental thickness taking into account associated VTEC values. Consequently, it is possible to include horizontal ionization gradients in the mapping procedure. Compared with the single thin-shell algorithm, the mapping function error is reduced by more than 50% under high as well as low solar activity conditions.

GPS Klobuchar Model

GPS offers a simple ionospheric TEC model for 50% corrections of single frequency measurements. This model or ionospheric correction algorithm (ICA), published by Klobuchar [6.79] provides a mean vertical delay at L1, for given geomagnetic location and local time. Here the diurnal variation of the vertical ionospheric delay is simply modeled by a half-cosine function with varying amplitude and period, depending on time and geomagnetic latitude (Fig. 6.12). During night time, the vertical ionospheric delay is fixed at a constant value of 5 ns (1.5 m at the L1 frequency).

The amplitude and period of the half-cosine form are centered at 14:00 local time. Accordingly, the time delay T_{ion} at L1 frequency is defined by the relation

$$T_{\text{ion}} = A_1 + A_2 \cos \left[\frac{2\pi(t_{\text{GPS}} - A_3)}{A_4} \right], \quad (6.100)$$

where A_1 is the constant night-time value (5 ns), A_2 is the amplitude, A_3 is a constant phase shift fixed at 14:00 local time and A_4 is the period of the cosine function.

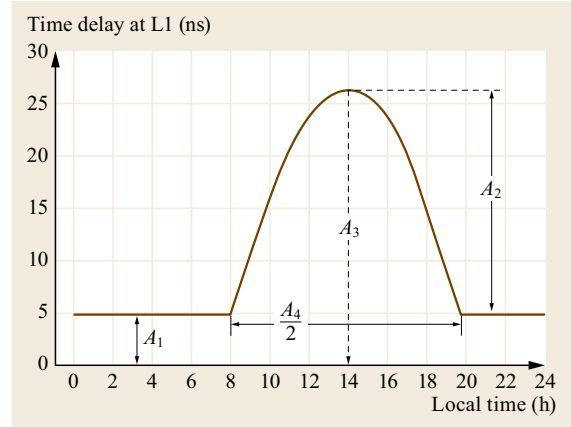


Fig. 6.12 Illustration of the Klobuchar GPS correction model

The model approach is realized by third-order polynomials for the amplitude A_2 and the period A_4 . The eight coefficients are updated daily by the GPS master control station, uploaded to the GPS satellites and transmitted back to the users via the navigation message. The amplitude and the period are defined by

$$A_2 = \sum_{n=0}^3 \alpha_n \phi_m \quad (6.101)$$

and

$$A_4 = \sum_{n=0}^3 \beta_n \phi_m. \quad (6.102)$$

Here, α_n and β_n denote satellite-transmitted coefficients of cubic polynomials for the amplitude of the vertical delay and the period of the model, whereas ϕ_m denotes the geomagnetic latitude of the Earth projection of the ionospheric piercing point. The mean ionospheric height is assumed to be 350 km.

Knowing receiver and satellite position, the ionospheric piercing point can immediately be determined to compute the associated vertical ionospheric delay using (6.100)–(6.102).

The obliquity factor $M(E)$ for transforming the vertical delay to the required slant delay is defined by a simple approach according to

$$M(E)_{\text{GPS}} = 1 + 16 (0.53 - E)^3, \quad (6.103)$$

where the elevation angle E is given in semicircles. A detailed description of the model can be found in [6.79].

It is worth to note that the ionospheric correction model of the Chinese Beidou Navigation Satellite System (BDS, previously known as COMPASS) is very

similar to the GPS model [6.80]. There are only a few modifications of the formulas given above. Thus, instead of the geomagnetic reference frame a geodetic reference frame is used in the COMPASS ionospheric model (CIM). Furthermore, instead of the mapping function approach $M(E)_{\text{GPS}}$ for GPS, the standard thin-shell mapping function (6.99) is used. The eight coefficients are updated every 2 h. A preliminary evaluation of CIM has indicated a similar over-all performance as obtained by the Klobuchar model – slightly better in the Northern Hemisphere and worse in the Southern Hemisphere [6.80].

The NeQuick Model

The NeQuick model is a three dimensional electron density model of the ionosphere/plasmasphere systems developed at the International Centre for Theoretical Physics (ICTP) Trieste, Italy, and at the University of Graz, Austria [6.81–83]. Compared with the international reference ionosphere (IRI) model [6.84] in favor of reduced computing time it is less complex and models only the electron density. Thus, it is tailored to numerically integrate the electron density along any GNSS satellite–receiver ray path to calculate TEC at global scale.

The vertical electron density profile is given by a sum of specific functions for the ionospheric layers E , F_1 , and F_2 . The topside ionosphere is described by a separate function whose scale height increases with height. The peak heights of different layers are also described by separate functions. The spatial and temporal behavior of key parameters such as $N_m F_2$ is deduced from monthly tables of Comité Consultatif International des Radiocommunications (CCIR) coefficients [6.85]. In addition to latitudinal and longitudinal dependence of these coefficients, a geomagnetic field dependence expressed by the modified dipole parameter $\text{modip } \mu$ was included (CCIR 1967). The parameter was introduced by Rawer in 1963 [6.86] according to

$$\tan(\mu) = \frac{I}{\sqrt{\cos \varphi}}, \quad (6.104)$$

where I is the magnetic inclination at 300 km and φ is the geographic latitude of the piercing point location. The CCIR coefficients are given for low and high solar activities characterized by the 12 months running mean of the sunspot number R at levels $R_{12} = 0$ and $R_{12} = 100$, respectively. To prepare coefficients for any other solar conditions the coefficients are linearly interpolated. Thus, the model is able to provide a global 3-D electron density distribution during a full solar cycle. Instead of sunspot number index R_{12} also the radio flux

index $F_{10.7}$ can be used via the relationship

$$F_{10.7} = 63.7 + 0.728 R_{12} + 0.00089 R_{12}^2. \quad (6.105)$$

A specific version, known as NeQuick-G, is used as the single frequency correction model in the European navigation satellite system Galileo [6.87]. Here, the external solar activity index R_{12} is replaced by an *effective ionization level* Az which is computed by the Galileo operation center in order to get the best representation of the current ionization level by the model. Az is valid for 24 h on a global scale and is defined by

$$Az(\mu) = a_0 + a_1 \mu + a_2 \mu^2, \quad (6.106)$$

where the coefficients a_0 , a_1 , and a_2 are optimized every 24 h and broadcasted to the user for computing TEC along ray paths that are used for positioning. Initial performance results obtained during the in-orbit validation (IOV) phase are reported in [6.88].

The NTCM Model

The development of a global TEC model was initiated in the Deutsches Zentrum für Luft- und Raumfahrt (DLR) to assist calibration, mapping, and forecasting in TEC monitoring procedures [6.89–91].

The TEC model NTCM-GL (Neustrelitz TEC Model-GLobal) described by Jakowski et al. [6.89] provides a multiplicative representation of temporal and spatial variations of global TEC for a full solar cycle. Essential dependencies from local time, season, geomagnetic field, and solar activity are treated in a similar manner as applied for regional models developed for TEC mapping [6.91]. New approaches were added for describing low latitude features such as the low latitude crest in an effective way. To keep the number of coefficients as small as possible, the terms describing the above mentioned dependencies are combined in a multiplicative way as shown in (6.107). The coefficients and data are related through a nonlinear system of equations given by

$$\text{VTEC}_{\text{NTCM-GL}} = F_{\text{LT}} F_{\text{seas}} F_{\text{mag}} F_{\text{crest}} F_{\text{sol}}. \quad (6.107)$$

The 12 coefficients are determined by an iterative nonlinear least-squares technique. The different terms describe specific approaches related to local time (F_{LT}), seasonal (F_{seas}) and geomagnetic field (F_{mag}) variation of TEC. The terms F_{crest} and F_{sol} describe longitudinal dependencies of low latitude crest parameters and solar activity dependency, respectively. The local time variation is basically described by diurnal, semi- and ter-diurnal harmonic functions. The seasonal variation includes an annual and semiannual harmonic function.

The latitudinal dependence is described by a dipole approach of the geomagnetic latitude and a special expression for the crests at both sides of the geomagnetic equator.

The solar activity level is quantified by the solar radio flux at 10.7 cm wavelength being a proxy of the ionizing radiation of the Sun in the EUV wavelength range. Further details of the model approach can be found in [6.89].

The model coefficients were determined by an iterative nonlinear least-squares technique applied to a long-term VTEC dataset from the Center for Orbit Determination in Europe (CODE) at the University of Berne as input. At CODE, the vertical TEC is modeled with a spherical harmonic expansion up to degree 15 and order 15 referring to a solar-geomagnetic reference frame [6.92, 93]. The two-hourly VTEC maps are derived from GPS data of the global network of the international GNSS service (IGS) [6.94]. The high-quality TEC dataset of the first NTCM-GL approach comprises data from more than 130 IGS stations from 1998–2007 over more than half a solar cycle. The nonlinear approach requires only 12 coefficients and the solar activity index $F_{10.7}$ as an external parameter for describing global TEC variations during a full solar cycle at all levels of solar activity. Due to the simplicity of the model approach, the model runs very fast and the implementation in operational systems is easy.

The IRI Model

The IRI is an empirical model widely used for different applications [6.84]. The development of IRI started

in 1968 co-sponsored by the Committee on Space Research (COSPAR) and the International Union of Radio Science (URSI).

IRI is a complex model that describes the electron concentration, electron temperature, ion temperature, and ion composition in the altitude range from 50 km to about 2000 km for a given location, time, and date. Due to the upper boundary limitation the vertical electron content can be computed only up to 2000 km height. The solar activity dependence is introduced by the solar radio flux $F_{10.7}$ or the sunspot number R_{12} . Concerning the description of a key parameter such as f_0F_2 , IRI can alternatively use CCIR coefficients adopted in 1967 as used also in NeQuick or a set of coefficients proposed by the URSI in 1989. Instead of using such coefficients, IRI can include also current observation values of key parameters such as N_mF_2 . The electron density distribution of the topside ionosphere is described by the NeQuick approach.

Newer versions of IRI also contain a storm model that improves the modeling results during ionospheric storms compared with the classical climatology approach [6.96]. The storm model is driven by the planetary geomagnetic K_p index. The IRI model is updated periodically and has evolved over a number of years as the result of the work of the international science community. The most recent version of IRI can be found in [6.97].

The major limitation of IRI previous versions in terms of electron density representation appears to be its electron distribution in the region above the peak of the F2 region. Due to the complex character of

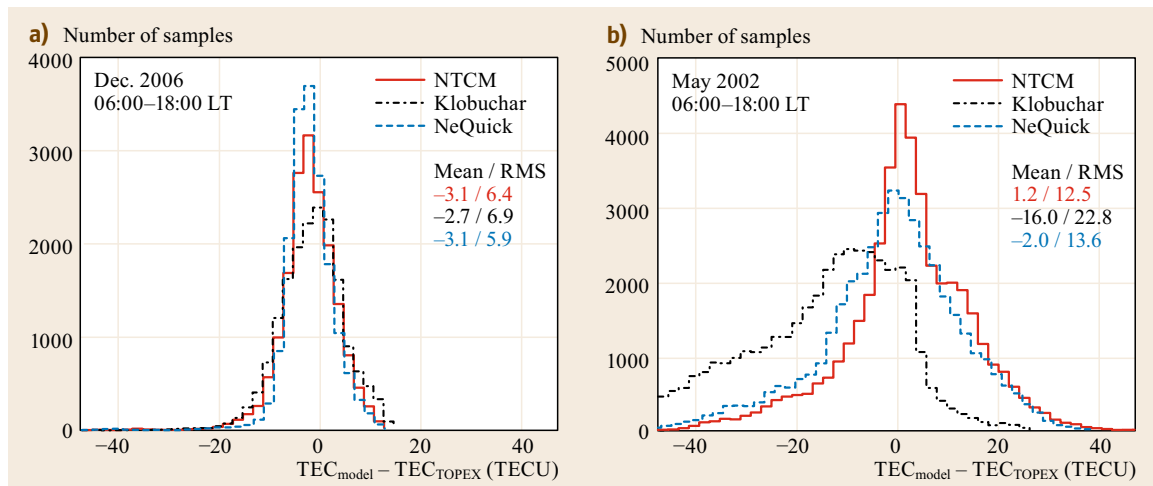


Fig. 6.13a,b Comparison of daytime ionospheric VTEC estimations derived from models: Klobuchar ICA-GPS, NeQuick and NTCM-GL with VTEC estimations from dual frequency satellite altimetry data at TOPEX/Poseidon. Selected data samples from December 2006 at low solar activity (a) and May 2002 at high solar activity (b). Mean deviations and RMS values are given in TECU (after [6.95])

IRI, it is well-suited for research and case studies. However, internal time consuming computations of various parameters such as ion composition and electron temperature are not needed for TEC computations. Furthermore, the limitation in height at 2000 km appears to be an obstacle in using IRI for GNSS-based TEC monitoring.

Comparison of Models

Whereas 3-D models allow estimating TEC by integrating the electron density distribution along the concrete ray path, 2-D TEC models require a mapping function, which is a principal source of errors in particular at low elevation angles as discussed in the previous section. Nevertheless, three dimensional electron density models such as IRI or NeQuick may also fail in estimating TEC even if the peak electron density is modeled very well [6.98].

The comparative analysis of three ionospheric correction models for single frequency transionospheric range errors clearly shows that TEC estimations of NeQuick and NTCM-GL are very similar (Fig. 6.13). This confirms earlier performance checks of NTCM-GL in comparison with NeQuick indicating that the 12 coefficient TEC model NTCM-GL achieves practically the same performance as the much more complex electron density model NeQuick [6.83]. Moreover, it has to be stated, that the Klobuchar or GPS model performance deviates from NeQuick and NTCM-GL significantly although the coefficients are regularly updated. The comparison is in particular worse at night time when TEC is fixed at 9.22 TECU (5 ns at L1 frequency) in the Klobuchar model.

6.3.5 Measurement-Based Ionosphere Correction

Single-Frequency Measurements

In single-frequency applications, ionospheric range errors can be estimated to a certain degree by utilizing model values [6.79], external TEC monitoring data [6.99] or the code-carrier divergence [6.100]. The latter solution is theoretically based on opposite signs of the first-order terms in the refractive indices for carrier and group phases as can be seen in (6.78) and (6.82). Utilizing this unique relationship, Yunck [6.100] proposed the *group and phase ionosphere correction* (GRAPHIC) in 1993 to mitigate the first-order ionospheric range error in single frequency GNSS signals. Computing the arithmetic mean of code and carrier phases, the ionospheric first-order term cancels out in a similar way as in dual frequency solutions. Compared with dual frequency code-measurements the GRAPHIC-method reduces the noise level by 50%. Although still rarely

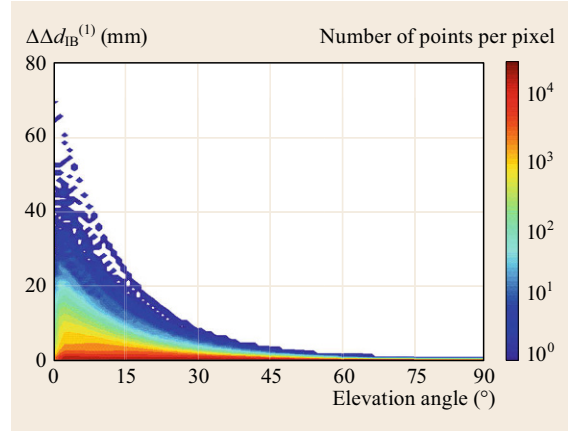


Fig. 6.14 Residual range error in the ionosphere-free linear combination of L1 and L2 frequencies due to different STEC obtained at L1 and L2 ray paths. First-order estimations have been carried out using electron density profiles derived from CHAMP radio occultation data in 2002 (after [6.104])

applied in terrestrial positioning [6.101], the method is wide-spread in spaceborne navigation [6.102]. The linear combination of code and carrier phases can even be used for ionospheric monitoring when code noise is low [6.103].

Dual-Frequency Measurements

To essentially reduce the ionospheric error in precise applications, dual-frequency measurements are made. Utilizing the dispersive nature of the ionosphere the first-order ionospheric error is mitigated in a linear combination of phase measurements at both frequencies according to

$$\varphi_{IF} = \frac{f_1^2}{f_1^2 - f_2^2} \varphi_1 - \frac{f_2^2}{f_1^2 - f_2^2} \varphi_2. \quad (6.108)$$

While the first-order error $d_1^{(1)}$ is cancelled out, some higher order terms including bending remain in the error budget. Following Hoque and Jakowski [6.53], we can write for the φ_{IF} linear combination

$$\varphi_{IF} = \rho + \Delta d_1^{(2)} + \Delta d_1^{(3)} - \Delta \Delta s_B + \Delta s_{\varphi_B}. \quad (6.109)$$

Besides the difference of higher order errors $\Delta d_1^{(2)}$ and $\Delta d_1^{(3)}$, the difference of the bending effect at both frequencies is considered.

In addition to former discussion the linear combination (6.108) requires to consider also the difference of phase error effect at different ray paths (Fig. 6.5) which we call Δs_{φ_B} . According to Hoque and Jakowski [6.53] different STEC values along different ray paths con-

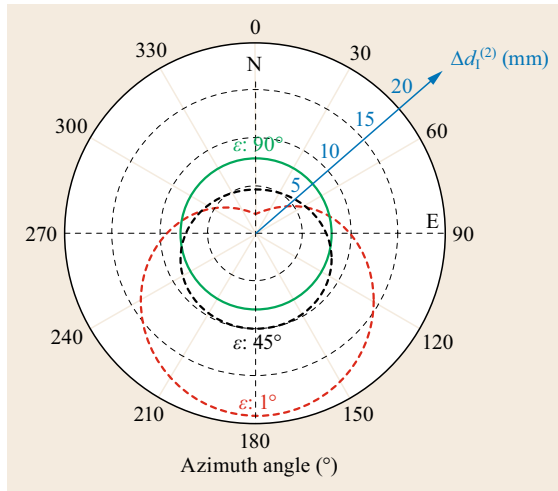


Fig. 6.15 Azimuth dependence of the second-order residual phase error for different elevation angles at the location 48°N, 15°E in Europe assuming a vertical TEC of 100 TECU (after [6.53])

tribute with a residual error of up to 6 cm as shown in Fig. 6.14.

A model approach for estimating TEC-related bending errors in the ionosphere-free linear combination of L1 and L2 measurements and a related statistics is published by Hoque and Jakowski [6.53]. The vertical structure of the ionosphere is modeled by a Chapman layer function as discussed in Sect. 6.3.1. The max-

imum error obtained in the statistical estimates was about 5 cm for 1° elevation.

Model computations have been performed also to mitigate the Faraday effect in dual frequency positioning [6.53, 105]. Figure 6.15 shows the typical asymmetry of the second-order residual phase range error in the ionosphere-free linear combination for L1 and L2 measurements. Since the geomagnetic field changes smoothly over a certain area like Germany or mid-Europe, model approaches with fixed geomagnetic geometry have been developed by Hoque and Jakowski [6.53, 106] to estimate the second-order residual carrier phase error for L1/L2 GPS frequencies. To characterize the ionosphere, only TEC is used as an input parameter. The achieved accuracy of the model approach is in the order of 2–3 mm for Germany and mid-Europe. The estimation may help at least to improve the phase ambiguity resolution in precise positioning.

A more detailed discussion of characteristic residual errors which remain in the so-called ionosphere-free linear combination is given by Hoque and Jakowski [6.53].

Acknowledgments. Norbert Jakowski would like to express his gratitude to his colleagues from the German Aerospace Center with whom he has worked over many years. In particular he thanks his colleague Dr. Mohammed Mainul Hoque for close cooperation for more than a decade.

References

- 6.1 D.J. Griffiths: *Introduction to Electrodynamics*, 4th edn. (Addison-Wesley, Boston 2012)
- 6.2 J.D. Jackson: *Classical Electrodynamics*, 3rd edn. (John Wiley, New York 1998)
- 6.3 H.J. Liebe: MPM – An atmospheric millimeter-wave propagation model, *Int. J. Infrared Millim. Wave* **10**(6), 631–650 (1989)
- 6.4 P. Debye: *Polar Molecules* (Dover, New York 1929)
- 6.5 K.G. Budden: *The Propagation of Radio Waves: The Theory of Radio Waves of Low Power in the Ionosphere and Magnetosphere*, 1st edn. (Cambridge Univ. Press, Cambridge 1985)
- 6.6 K. Davies: *Ionospheric Radio* (Peter Peregrinus, London 1990)
- 6.7 L. Essen, K.D. Froome: Dielectric constant and refractive index of air and its principal constituents at 24,000 Mc/s, *Nature* **167**, 512–513 (1951)
- 6.8 J.C. Owens: Optical refractive index of air: Dependence on pressure, temperature and composition, *Appl. Opt.* **6**(1), 51–59 (1967)
- 6.9 J.M. Rüeger: Refractive index formula for radio waves, *Proc. XXII FIG Int. Congr.*, Washington (FIG, Copenhagen 2002) pp. 1–13
- 6.10 J. Böhm, H. Schuh: *Atmospheric Effects in Space Geodesy* (Springer, Berlin 2013)
- 6.11 J. Saastamoinen: Atmospheric correction for the troposphere and stratosphere in radio ranging satellites. In: *The Use of Artificial Satellites for Geodesy*, ed. by S.W. Henriksen, A. Mancini, B.H. Chovitz (AGU, Washington 1972) pp. 247–251
- 6.12 H. Berg: *Allgemeine Meteorologie* (Dümmler, Berlin 1948)
- 6.13 B. Hofmann-Wellenhof, H. Moritz: *Physical Geodesy* (Springer, Berlin 2006)
- 6.14 H.S. Hopfield: Two-quartic tropospheric refractivity profile for correcting satellite data, *J. Geophys. Res.* **74**(18), 4487–4499 (1969)
- 6.15 R.F. Leandro, M.C. Santos, R.B. Langley: UNB neutral atmosphere models: Development and performance, *Proc. ION NTM 2006*, Monterey (ION, Virginia 2006) pp. 564–573
- 6.16 J. Boehm, R. Heinkelmann, H. Schuh: Short note: A global model of pressure and temperature for geodetic applications, *J. Geodesy* **81**(10), 679–683

- (2007)
- 6.17 K. Lagler, M. Schindelegger, J. Boehm, H. Krasna, T. Nilsson: GPT2: Empirical slant delay model for radio space geodetic techniques, *Geophys. Res. Lett.* **40**(6), 1069–1073 (2013)
 - 6.18 G. Petit, B. Luzum: *IERS Conventions (2010)* (Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt 2010), IERS Technical Note No. 36
 - 6.19 United States Committee on Extension to the Standard Atmosphere: *US Standard Atmosphere Supplements 1966* (US Govt. Print. Off., Washington 1966)
 - 6.20 J.L. Davis, T.A. Herring, I.I. Shapiro, A.E.E. Rogers, G. Elgered: Geodesy by radio interferometry: Effects of atmospheric modeling errors on estimates of baseline length, *Radio Sci.* **20**, 1593–1607 (1985)
 - 6.21 V.B. Mendes: Modeling the Neutral-Atmosphere Propagation Delay in Radiometric Space Techniques, Ph.D. Thesis (Univ. New Brunswick, Fredericton 1999)
 - 6.22 C.C. Chao: *A Model for Tropospheric Calibration from Daily Surface and Radiosonde Balloon Measurement*, Tech. Mem. 391–350 (Jet Propulsion Laboratory, Pasadena 1972) pp. 67–73
 - 6.23 C.C. Chao: *New Tropospheric Range Corrections with Seasonal Adjustment*, DSN Progr. Rep., JPL Report No. 32–1526, Vol. (Jet Propulsion Laboratory, Pasadena 1971) pp. 67–73
 - 6.24 C.C. Chao: *A New Method to Predict Wet Zenith Range Refraction from Surface Measurements of Meteorological Parameters*, DSN Progr. Rep. No. 32–1526 (Jet Propulsion Laboratory, Pasadena 1973) pp. 33–41
 - 6.25 H.S. Hopfield: The effect of tropospheric refraction on the Doppler shift of a satellite signal, *J. Geophys. Res.* **68**(18), 5157–5168 (1961)
 - 6.26 H.S. Hopfield: Tropospheric effect on electromagnetically measured range: Prediction from surface weather data, *Radio Sci.* **6**(3), 357–367 (1972)
 - 6.27 H.S. Hopfield: *Tropospheric Effects on Signals at Very Low Elevation Angles* (Appl. Phys. Lab., John Hopkins Univ., Laurel 1976), Tech. Memo. TG1291
 - 6.28 H.S. Hopfield: Improvements in the tropospheric refraction correction for range measurement, *Philos. Trans. R. Soc. Lond.* **294**(1410), 341–352 (1979)
 - 6.29 J.W. Marini: Correction of satellite tracking data for an arbitrary tropospheric profile, *Radio Sci.* **7**(2), 223–231 (1972)
 - 6.30 T.A. Herring: Modelling atmospheric delay in the analysis of space geodetic data. In: *Symposium on Refraction of Transatmospheric Signals in Geodesy*, Publications on Geodesy, No. 36, ed. by J.C. de Munck, T.A.T. Spoelstra (Netherlands Geodetic Commission, Delft 1992) pp. 157–164
 - 6.31 A.E. Niell: Global mapping functions for the atmosphere delay at radio wavelengths, *J. Geophys. Res.* **101**(B2), 3227–3246 (1996)
 - 6.32 L.P. Gradinarsky, J.M. Johansson, G. Elgered, P. Jarlemar: GPS site testing at Chajnantor in Chile, *Phys. Chem. Earth* **26**(6–8), 421–426 (2001)
 - 6.33 C. Rocken, S. Sokolovskiy, J.M. Johnson, D. Hunt: Improved mapping of tropospheric delays, *J. Atmos. Ocean. Technol.* **18**, 1205–1213 (2001)
 - 6.34 A.E. Niell: Improved atmospheric mapping functions for VLBI and GPS, *Earth Planets Space* **52**, 699–702 (2000)
 - 6.35 A.E. Niell: Global mapping functions for the atmosphere delay at radio wavelengths, *Phys. Chem. Earth* **26**(6–8), 475–480 (2001)
 - 6.36 J. Boehm, H. Schuh: Vienna mapping functions in VLBI analyses, *Geophys. Res. Lett.* **31**(L01603), 1–4 (2004)
 - 6.37 J. Boehm, B. Werl, H. Schuh: Troposphere mapping functions for GPS and very long baseline interferometry from European centre for medium-range weather forecasts operational analysis data, *J. Geophys. Res.* **111**(B02406), 1–9 (2006)
 - 6.38 Vienna University of Technology, GGOS Atmosphere: Atmosphere Delays (Vienna Univ. Technology, Vienna 2014) <http://ggosatm.hg.tuwien.ac.at/delay.html>
 - 6.39 L. Urquhart, M. Santos, F. Nievinski, J. Böhm: Generation and assessment of VMF1-type grids using North-American numerical weather models. In: *Earth on the Edge: Science for a Sustainable Planet*, ed. by C. Rizos, P. Willis (Springer, Berlin 2014) pp. 3–9
 - 6.40 *University of New Brunswick Vienna Mapping Function Service* (Univ. New-Brunswick, Fredericton) <http://unb-vmf1.gge.unb.ca/>
 - 6.41 J. Böhm, A. Niell, P. Tregoning, H. Schuh: Global mapping function (GMF): A new empirical mapping function based on data from numerical weather model data, *Geophys. Res. Lett.* **33**(L07304), 1–4 (2006)
 - 6.42 P. Gegout, R. Biancale, L. Soudarin: Adaptive mapping functions to the azimuthal anisotropy of the neutral atmosphere, *J. Geodesy* **85**(6–8), 661–677 (2011)
 - 6.43 Th. Hobiger, R. Ichikawa, T. Takasu, Y. Koyama, T. Kondo: Ray-traced troposphere slant delays for precise point positioning, *Earth Planets Space* **60**(5), 1–4 (2008)
 - 6.44 F.G. Nievinski: Ray-Tracing Options to Mitigate the Neutral Atmosphere Delay in GPS, Ph.D. Thesis (Univ. New Brunswick, Fredericton 2008)
 - 6.45 V. Nafisi, M. Madzak, J. Böhm, A.A. Ardalan, H. Schuh: Ray-traced tropospheric delays in VLBI analysis, *Radio Sci.* **47**(RS2020), 1–17 (2012)
 - 6.46 D.S. MacMillan: Atmospheric gradients from very long baseline interferometry observations, *Geophys. Res. Lett.* **22**(9), 1041–1044 (1995)
 - 6.47 G. Chen, T.A. Herring: Effects of atmospheric azimuthal asymmetry on the analysis of space geodetic data, *J. Geophys. Res. Solid Earth* **102**(B9), 20489–20502 (1997)
 - 6.48 S. Chapman: The absorption and dissociative or ionizing effect of monochromatic radiation in an atmosphere on a rotating earth, *Proc. Phys. Soc.* **43**, 1047–1055 (1931)
 - 6.49 K. Rawer: *Wave Propagation in the Ionosphere* (Kluwer, Dordrecht 1993)
 - 6.50 G.K. Hartmann, R. Leitinger: Range errors due to ionospheric and tropospheric effects for signal fre-

- quencies above 100 MHz, Bull. Géodésique **58**(2), 109–136 (1984)
- 6.51 S. Bassiri, G.A. Hajj: Higher-order ionospheric effects on the global positioning system observables and means of modeling them, Manuscripta Geodaetica **18**(6), 280–289 (1993)
- 6.52 M.M. Hoque, N. Jakowski: Higher-order ionospheric effects in precise GNSS positioning, J. Geodesy **81**(4), 280–289 (2006)
- 6.53 M.M. Hoque, N. Jakowski: Estimate of higher order ionospheric errors in GNSS positioning, Radio Sci. **43**(RS5008), 1–15 (2008)
- 6.54 B.W. Parkinson, S.W. Gilbert: NAVSTAR: Global positioning system – Ten years later, Proc. IEEE **71**(10), 1177–1186 (1983)
- 6.55 S. Kedar, G. Hajj, B. Wilson, M. Heflin: The effect of the second order GPS ionospheric correction on receiver position, Geophys. Res. Lett. **30**(16), 1829 (2003)
- 6.56 M. Hernandez-Pajares, J.M. Juan, J.M. Sanz, R. Orus: Second order ionospheric term in GPS: Implementation and impact on geodetic estimates, J. Geophys. Res. **112**(B08417), 1–16 (2007)
- 6.57 R. Leitinger, E. Putz: Ionospheric refraction errors and observables. In: *Atmospheric Effects on the Geodetic Space Measurements*, Monograph 12, ed. by F.K. Brunner (School of Surveying, UNSW, Kensington 1988) pp. 81–102
- 6.58 F.K. Brunner, M. Gu: An improved model for the dual frequency ionospheric correction of GPS observations, Manuscripta Geodaetica **16**(3), 205–214 (1991)
- 6.59 N. Jakowski, F. Porsch, G. Mayer: Ionosphere-induced-ray-path bending effects in precise satellite positioning systems, Z. Satell. Position. Navig. Kommun. **3**(1), 6–13 (1994)
- 6.60 M.M. Hoque, N. Jakowski: Higher order ionospheric propagation effects on GPS radio occultation signals, Adv. Space Res. **46**(2), 162–173 (2010)
- 6.61 M.M. Hoque, N. Jakowski: Ionospheric bending correction for GNSS radio occultation signals, Radio Sci. **46**(RS0D06), 1–9 (2011)
- 6.62 R.D. Hunsucker: *Radio Techniques for Probing the Terrestrial Ionosphere* (Springer, Berlin 1991)
- 6.63 L. Barclay (Ed.): *Propagation of Radio Waves*, 2nd edn. (IET, London 2003)
- 6.64 P.M. Kintner, B.M. Ledvina: The ionosphere, radio navigation, and global navigation satellite systems, Adv. Space Res. **32**(5), 788–811 (2005)
- 6.65 M.C. Kelley: *The Earth's Ionosphere – Plasma Physics and Electrodynamics*, 2nd edn. (Elsevier, Amsterdam 2009)
- 6.66 L. Alfonsi, G. De Franceschi, V. Romano, A. Bourdillon, M. Le Huy: GPS scintillations and TEC gradients at equatorial latitudes on April 2006, Adv. Space Res. **47**(10), 1750–1757 (2011)
- 6.67 A.M. Smith, C.N. Mitchell, R.J. Watson, R.W. Meggs, P.M. Kintner, K. Kauristie, F. Honary: GPS scintillation in the high arctic associated with an auroral arc, Space Weather **6**(S03D01), 1–7 (2008)
- 6.68 S. Fukao, T. Yokoyama, T. Tayama, M. Yamamoto, T. Maruyama, S. Saito: Eastward traverse of equatorial plasma plumes observed with the equatorial atmosphere radar in Indonesia, Ann. Geophysicae. **24**(5), 1411–1418 (2006)
- 6.69 M. Nishioka, A. Saito, T. Tsugawa: Occurrence characteristics of plasma bubble derived from global ground-based GPS receiver networks, J. Geophys. Res. **113**(A05301), 1–12 (2008)
- 6.70 S. Basu, E. MacKenzie, S. Basu: Ionospheric constraints on VHF/UHF communication links during solar maximum and minimum period, Radio Sci. **23**(3), 363–378 (1988)
- 6.71 J.A. Secan, R.M. Bussey, E.J. Fremouw, S. Basu: High-latitude upgrade to the wideband ionospheric scintillation model, Radio Sci. **32**(4), 1567–1574 (1997)
- 6.72 Y. Bénéguet: Global ionospheric propagation model (GIM): A propagation model for scintillations of transmitted signals, Radio Sci. **32**(3), 1–13 (2002)
- 6.73 Y. Bénéguet, P. Hamel: A global ionosphere scintillation propagation model for equatorial regions, J. Space Weather Space Clim. **1**(A04), 1–8 (2011)
- 6.74 A. Komjathy, L. Sparks, A.J. Mannucci, A. Coster: The ionospheric impact of the October 2003 storm event on WAAS, Proc. ION GNSS 2004, Long Beach (ION, Virginia 2004) pp. 1298–1307
- 6.75 T. Sakai, T. Yoshihara, S. Saito, K. Matsunaga, K. Hoshinoo, T. Walter: Modeling vertical structure of ionosphere for SBAS, Proc. ION GNSS 2009, Savannah (ION, Virginia 2009) pp. 1257–1267
- 6.76 A.J. Mannucci, B. Iijima, L. Sparks, X. Pi, B. Wilson, B.U. Lindqwister: Assessment of global TEC mapping using a three-dimensional electron density model, J. Atmos. Sol. Terr. Phys. **61**, 1227–1236 (1999)
- 6.77 M. Hernandez-Pajares, J.M. Juan, J. Sanz, M. Garcia-Fernandez: Towards a more realistic ionospheric mapping function, Proc. XXVIII URSI Gen. Assembly, Delhi (URSI, Ghent 2005) pp. 1–4
- 6.78 M.M. Hoque-Pajares, N. Jakowski: Mitigation of ionospheric mapping function error, Proc. ION GNSS 2013, Nashville (ION, Virginia 2013) pp. 1848–1855
- 6.79 J.A. Klobuchar: Ionospheric time-delay algorithm for single-frequency GPS users, IEEE Trans. Aerosp. Electron. Syst. **23**(3), 325–331 (1987)
- 6.80 X. Wu, X. Hu, G. Wang, H. Zhong, C. Tang: Evaluation of COMPASS ionospheric model in GNSS positioning, Adv. Space Res. **51**(6), 959–968 (2013)
- 6.81 G. Hochegger, B. Nava, S. Radicella, R. Leitinger: A family of ionospheric models for different uses, Phys. Chem. Earth **25**(4), 307–310 (2000), Part C
- 6.82 S.M. Radicella, R. Leitinger: The evolution of the DGR approach to model electron density profiles, Adv. Space Res. **27**(1), 35–40 (2001)
- 6.83 B. Nava, P. Coisson, S.M. Radicella: A new version of the NeQuick ionosphere electron density model, J. Atmos. Sol.-Terr. Phys. **70**(15), 1856–1862 (2008)
- 6.84 D. Bilitza: International reference ionosphere, Radio Sci. **36**(2), 261–275 (2001)
- 6.85 W.B. Jones, R.M. Gallet: The representation of diurnal and geographical variations of ionospheric data by numerical methods, ITU Telecomm. J. **29**(5), 129–149 (1962)

- 6.86 K. Rawer: *Meteorological and Astronomical Influences on Radio Wave Propagation* (Academic, New York 1963) pp. 221–250
- 6.87 European GNSS (Galileo) Open Service: Ionospheric correction algorithm for Galileo single frequency users, Iss. 1.1, Feb. 2015 (EU 2015), doi:10.2873/723786
- 6.88 R. Orus-Pérez, R. Prieto-Cerdeira, B. Arbesser-Rastburg: The Galileo single-frequency ionospheric correction and positioning observed near the solar cycle 24 maximum, Proc. 4th Int. Coll. Sci. Fundam. Asp. the Galileo Prog., Prague (ESA, Noordwijk 2013)
- 6.89 N. Jakowski, M.M. Hoque, C. Mayer: A new global TEC model for estimating transionospheric radio wave propagation errors, J. Geodesy **85**(12), 965–974 (2011)
- 6.90 N. Jakowski, C. Mayer, M.M. Hoque, V. Wilken: TEC models and their use in ionosphere monitoring, Radio Sci. **46**(RS0D18), 1–11 (2011)
- 6.91 N. Jakowski, E. Sardon, S. Schlueter: GPS-based TEC observations in comparison with IRI95 and the European TEC model NTCM2, Adv. Space Res. **22**(6), 803–806 (1998)
- 6.92 S. Schaer: Mapping and Predicting the Earth's Ionosphere Using the Global Positioning System, Ph.D. Thesis (Astronomical Institute, Univ. Bern, Berne 1999)
- 6.93 M. Hernández-Pajares, J.M. Juan, J. Sanz, R. Orus, A. García-Rigo, J. Feltens, A. Komjathy, S.C. Schaer, A. Krankowski: The IGS VTEC maps: A reliable source of ionospheric information since 1998, J. Geodesy **83**(3/4), 263–275 (2009)
- 6.94 J.M. Dow, R.E. Neilan, C. Rizos: The international GNSS service in a changing landscape of global navigation satellite systems, J. Geodesy **83**(3/4), 191–198 (2009)
- 6.95 N. Jakowski, M.M. Hoque: Ionospheric range error correction models, Proc. Int. Conf. Localiz. GNSS (ICL-GNSS), Starnberg (2012) pp. 1–6
- 6.96 E.A. Araujo-Pradere, T.J. Fuller-Rowell, D. Bilitza: Validation of the STORM response in IRI2000, J. Geophys. Res. Space Phys. **108**(A3), 1–10 (2003)
- 6.97 D. Bilitza: *International Reference Ionosphere* (NASA GSFC, Greenbelt) <http://iri.gsfc.nasa.gov/>
- 6.98 P. Coisson, S.M. Radicella, R. Leitinger, B. Nava: Topside electron density in IRI and NeQuick: Features and limitations, Adv. Space Res. **37**(5), 937–942 (2006)
- 6.99 A.Q. Le, C.C.J.M. Tiberius, H. van der Marel, N. Jakowski: Use of global and regional ionosphere maps for single-frequency precise point positioning. In: *Observing our Changing Earth*, ed. by M.G. Sideris (Springer, Berlin 2008) pp. 759–769
- 6.100 T.P. Yunck: Coping with the atmosphere and ionosphere in precise satellite and ground positioning. In: *Environmental Effects on Spacecraft Positioning and Trajectories*, ed. by A.V. Jones (AGU, Washington 1992) pp. 1–16
- 6.101 T. Schüler, H. Diessongo, Y. Poku-Gyamfi: Precise ionosphere-free single-frequency GNSS positioning, GPS Solutions **15**(2), 139–147 (2011)
- 6.102 O. Montenbruck, T.V. Helleputte, R. Kroes, E. Gill: Reduced dynamic orbit determination using GPS code and carrier measurements, Aerosp. Sci. Technol. **9**(3), 261–271 (2005)
- 6.103 T. Schüler, O. Abel Oladipo: Single-frequency GNSS retrieval of vertical total electron content (VTEC) with GPS L1 and Galileo E5 measurements, J. Space Weather Space Clim. **3**(A11), 1–8 (2013)
- 6.104 N. Jakowski: Ionospheric GPS radio occultation measurements on board CHAMP, GPS Solutions **9**(2), 88–95 (2005)
- 6.105 S. Datta-Barua, T. Walter, J. Blanch, P. Enge: Bounding higher-order ionosphere errors for the dual-frequency GPS user, Radio Sci. **43**(RS5010), 1–15 (2008)
- 6.106 M. Hoque, N. Jakowski: Mitigation of higher order ionospheric effects on GNSS users in Europe, GPS Solutions **12**(2), 87–97 (2007)

Part B Satellite

Part B Satellite Navigation Systems

7 The Global Positioning System (GPS)

Christopher J. Hegarty, Bedford, USA

8 GLONASS

Sergey Revnivkykh, Moscow,

Russian Federation

Alexey Bolkunov, Korolyov,

Russian Federation

Alexander Serdyukov, Korolyov,

Russian Federation

Oliver Montenbruck, Wessling, Germany

9 Galileo

Marco Falcone, Noordwijk,

The Netherlands

Jörg Hahn, Noordwijk, The Netherlands

Thomas Burger, Noordwijk,

The Netherlands

10 Chinese Navigation Satellite Systems

Yuanxi Yang, Beijing, China

Jing Tang, Beijing, China

Oliver Montenbruck, Wessling, Germany

11 Regional Systems

Satoshi Kogure, Tokyo, Japan

A.S. Ganeshan, Bangalore, India

Oliver Montenbruck, Wessling, Germany

12 Satellite Based Augmentation Systems

Todd Walter, Stanford, USA

7. The Global Positioning System (GPS)

Christopher J. Hegarty

This chapter presents an overview of the US Global Positioning System (GPS), which became the first operational global navigation satellite system (GNSS) core constellation when it was declared fully operational in 1995. First, the space segment is described, including key characteristics of the different satellite types. Then, an overview of the control segment is given, including its operations and evolution of capabilities. This is followed by an overview of the GPS signals, current and future, as well as a description of the navigation data content. Then, the time and coordinate systems used by GPS are described. The chapter is concluded with a brief description of services and performance.

7.1	Space Segment	197
7.1.1	Constellation Design and Management ..	197
7.1.2	GPS Satellites.....	199
7.2	Control Segment	203
7.2.1	Overview.....	203
7.2.2	Evolution of Capabilities.....	204
7.2.3	Operations	204
7.3	Navigation Signals	205
7.3.1	Legacy	205
7.3.2	Modernized Signals.....	206
7.3.3	Power Levels	209
7.4	Navigation Data and Algorithms	210
7.4.1	Legacy Navigation (LNAV) Data Overview ..	210
7.4.2	LNAV Error Detection Encoding	211
7.4.3	LNAV Data Content and Related Algorithms.....	211
7.4.4	Civil Navigation (CNAV) and Civil Navigation-2 (CNAV-2) Data	215
7.5	Time System and Geodesy	216
7.6	Services and Performance	216
	References	217

The Global Positioning System (GPS) [7.1–4] is a satellite navigation system operated by the United States. The system consists of a constellation of nominally 24 satellites in medium altitude earth orbit (MEO), as well as a worldwide ground network to monitor and control the satellites. The GPS program began in the early 1970s and the system was declared fully opera-

tional in 1995. Internationally, the GPS constellation is considered to be just one component within the global collection of navigation satellites that is referred to as the global navigation satellite system (GNSS). This chapter provides an overview of GPS, including its space and control segments (CS), signals, services, and performance.

7.1 Space Segment

7.1.1 Constellation Design and Management

The GPS constellation nominally consists of 24 satellites in circular orbits with an orbital radius of 26 559 km [7.5] (Table 7.1). The satellite orbits are inclined 55° with respect to the equatorial plane. Four satellites are contained in each of six orbital planes,

which are equally spaced with respect to their orientation around the Earth's spin axis. The six orbital planes are identified by letter designators, from A to F. The nominal 24 satellite locations at a specified epoch are referred to as *slots*, and are designated by a letter-number combination, for example, A1 for the first satellite slot within the A plane. The nominal slot

Table 7.1 Nominal GPS constellation parameters

Parameter	Value
Number of operational satellites	24
Number of orbital planes	6
Number of satellites in a plane	4
Orbit type	Near circular
Eccentricity	$e < 0.02$
Inclination	$i = 55^\circ$
Nominal orbital altitude	$h = 20\,180\text{ km}$
Period of revolution	$T = 11\text{ h } 58\text{ m}$
Long. of asc. node between planes	$\Delta\Omega = 60^\circ$
Ground track repeat cycle	2 orbit/ $1_{\text{sid}}^{\text{d}}$

locations for $00^{\text{h}}\,00^{\text{m}}\,00^{\text{s}}$ coordinated universal time (UTC), 1 July, 1993, are provided in Fig. 7.1 [7.5]. The Greenwich hour angle for this epoch is $18^{\text{h}}\,36^{\text{m}}\,14.4^{\text{s}}$.

From Fig. 7.1, it may be noted that the four slots within each orbital plane are asymmetrically spaced. This design was determined to be robust against probable satellite failures [7.6]. The orbital altitude was chosen, in part, to support early system testing when the constellation was only partially populated. The nominal altitude provides an orbital period of one-half a sidereal day (approximately $11^{\text{h}}\,58^{\text{m}}$) so that the satellite ground tracks repeat daily (Fig. 7.2). Although repeating ground tracks are convenient for planning certain GPS applications, they result in resonant forces on each GPS satellite due to the Earth's nonuniform gravitational field, which in turn results in the need for more frequent satellite station-keeping maneuvers. The current constellation design was also influenced by many historical constraints that no longer apply, including early plans to launch the GPS satellites using the space shuttle that were abandoned after the Challenger disaster.

Table 7.2 Expandable slots in the baseline GPS 24-satellite constellation

Expandable slot		Right ascension of ascending node (RAAN)	Arg. of latitude
B1 expands to:	B1F	332.847°	94.916°
	B1A	332.847°	66.356°
D2 expands to:	D2F	92.847°	282.676°
	D2A	92.847°	257.976°
F2 expands to:	F2F	212.847°	0.456°
	F2A	212.847°	334.016°

In recent years, the constellation has been overpopulated with up to 31 operational satellites. The first 3 satellites beyond 24 are placed into *expandable slots* within the baseline 24-satellite constellation [7.5]. Each of the three slots B1, D2, and F2 may be split into two slots as shown in Table 7.2 to accommodate up to 27 total satellites in the constellation. *Surplus* satellites (operational satellites beyond the 27th) are typically placed in locations adjacent to satellites that are expected to require replacement the soonest.

As with any satellite constellation, occasional station-keeping maneuvers are required to keep the GPS satellites close to their nominal positions (slots). GPS satellite maneuvers are performed as necessary (typically once every 1–2 years per satellite) with the goal of keeping for each satellite the eccentricity within the range of 0–0.02, the inclination within the range of 52° – 58° , and the argument of latitude spacing within 4° of nominal values [7.5]. When a GPS satellite has reached end of life, its navigation signals are switched off and it is boosted by around 500 km in altitude to a disposal orbit [7.7].

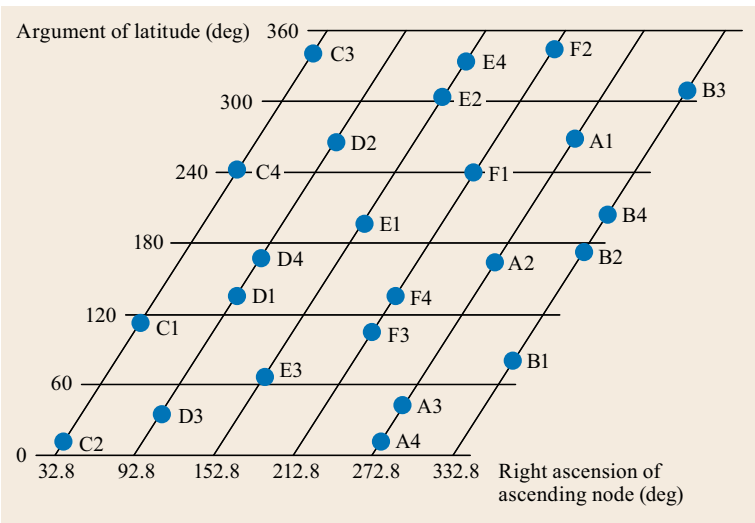


Fig. 7.1 Nominal GPS 24-satellite constellation for 1 July 1993 (after [7.5])

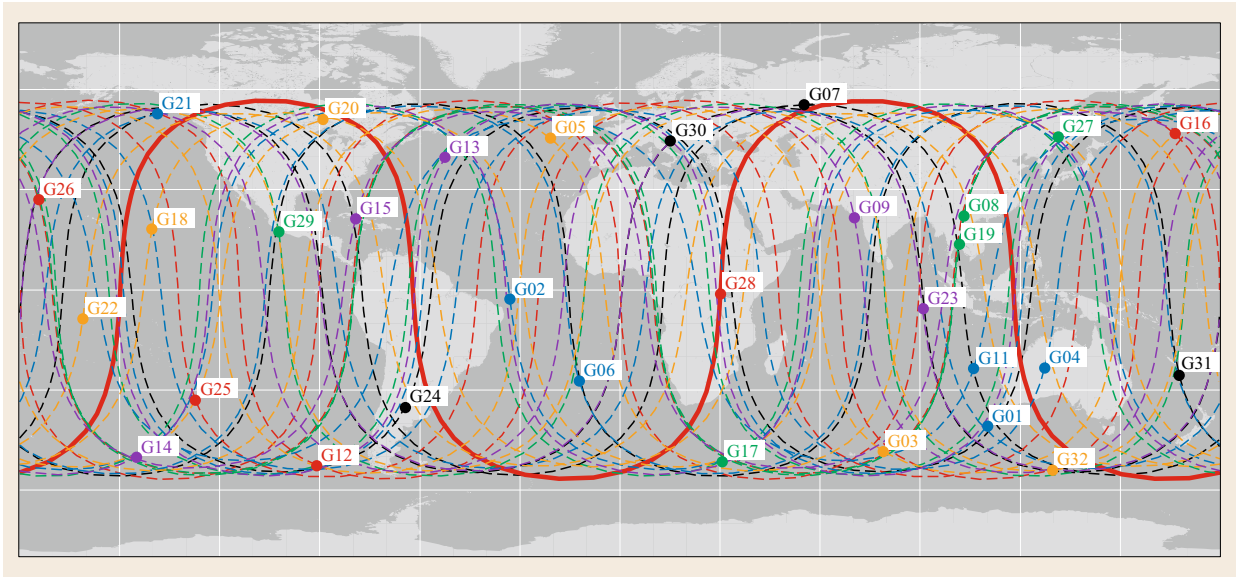


Fig. 7.2 Ground tracks of the GPS satellites for September 1, 2015, 00:00–24:00 UTC. Markers indicate the positions at the initial midnight epoch and different colors are used to distinguish the six orbital planes (black: A, red: B, green: C, blue: D, orange: E, orchid: F)

7.1.2 GPS Satellites

From 1978 until the present time, 67 GPS satellites were successfully launched into orbit out of which 31 are currently operational. Redundant atomic clocks, rubidium and/or cesium, are key components of each satellite so that signals that are precisely synchronized to a common timescale can be broadcast. The capabil-

ities of the satellites have increased over time, as has their size, weight, and cost. Some key characteristics of each satellite type are listed in Table 7.3.

Block I Satellites

A contract to build and launch a set of test satellites for GPS, referred to as the Block I space vehicles (SVs), was awarded to Rockwell International in Au-

Table 7.3 GPS satellites overview

Parameter	Block I	Block II/IIA	Block IIR/IIR-M	Block IIF	GPS III
First launch	1978	1989	1997	2010	2017 (planned)
Manufacturer	Rockwell International	Rockwell International	General Electric's Astro Space Division (now Lockheed Martin)	Rockwell International (now Boeing)	Lockheed Martin
Design life-time (years)	5	7.5	7.5	12	15
Mass (kg)	450	> 850	1080	1630	2200
System power (W)	400	700	1140	2610	4480
Solar array size (m ²)	5	7.2	13.6	22.2	28.5
Navigation payload					
Clocks	Rb, Cs	Cs, Rb	Rb	Cs, Rb	Rb
Clock stability (daily)	$2 \cdot 10^{-13}$, $1 \cdot 10^{-13}$	$1 \cdot 10^{-13}$, $5 \cdot 10^{-14}$	$1 \cdot 10^{-14}$	$1 \cdot 10^{-13}$, $0.5\text{--}1 \cdot 10^{-14}$	$5 \cdot 10^{-14}$
Signals	L1, L2	L1, L2	L1, L2	L1, L2, L5	L1, L2, L5
Cross link		×	×	×	×
Laser reflector	—	×	—	—	×
		(space vehicle number (SVN) 35, 36)			(later satellites)

gust 1974 [7.8]. Eleven Block I satellites, designated as GPS space vehicle numbers (SVN) 1–11, were launched from Vandenberg Air Force base (approximately 230 km northwest of Los Angeles) on refurbished Atlas-E/F intercontinental ballistic missiles from February 1978 to October 1985. All but one made it successfully into orbit. The exception was SVN 7, which was destroyed in a launch vehicle failure in December 1981. All of the Block I satellites carried three rubidium clocks, and the last eight additionally carried one cesium clock [7.9]. The Block I satellites had an on-orbit mass of approximately 450 kg and spanned 5.3 m from end to end of the deployed solar arrays. The Block I satellite power system included two solar arrays with a combined area of 5 m² to provide approximately 400 W of power and nickel cadmium (NiCd) batteries for power storage. The last six Block I and all subsequent GPS satellite blocks carry a secondary payload to detect nuclear explosions in the Earth's atmosphere and near space. The Block I satellites had a design life of 5 years, but some were in service for over 10 years. The last operational Block I satellite was decommissioned in late 1995.

Block II/IIA Satellites

In 1983, Rockwell International was awarded a contract to build and launch 28 operational GPS satellites, referred to as Block II. In March 1984, a decision was made to modify the 10th and successive Block II satellites to allow extended operations of up to 180 days

without ground contact, in addition to several other new capabilities. The modified vehicles are referred to as Block IIA. The 9 Block II and 19 Block IIA satellites were launched between February 1989 and November 1997. All of these satellites were designed to carry two rubidium and one cesium clock. The design life was 7.5 years. As of today (2015), only three Block IIA satellites remain in operation. The remaining Block II/IIA satellites have been decommissioned.

A Block IIA satellite is shown in Fig. 7.3. Two solar arrays are visible on either side of the main body of the satellite. The solar arrays cover an area of 7.2 m² and are part of a 700 W power system that also includes NiCd batteries for energy storage. The wingspan of the satellite is approximately the same as for the Block I satellites, 5.3 m. The on-orbit mass is 990 kg, which is significantly higher than the Block I satellites and slightly higher than the 850 kg Block II satellites. An array antenna with 12 helical elements arranged in two concentric rings (8 in the outer ring and 4 in the inner ring) is visible on the main body of the spacecraft. This array antenna is used to broadcast the L-band navigation signals from the spacecraft to the Earth. The L-band antenna is boresighted toward the center of the Earth. It is designed to direct most of the radiated power toward the surface of the Earth where most GPS users are expected to be. The antenna is broadband and has a peak gain of 13.2 dBi at 1575 MHz. The antenna gain pattern was optimized to provide near uniform received power levels over the surface of the Earth [7.10].

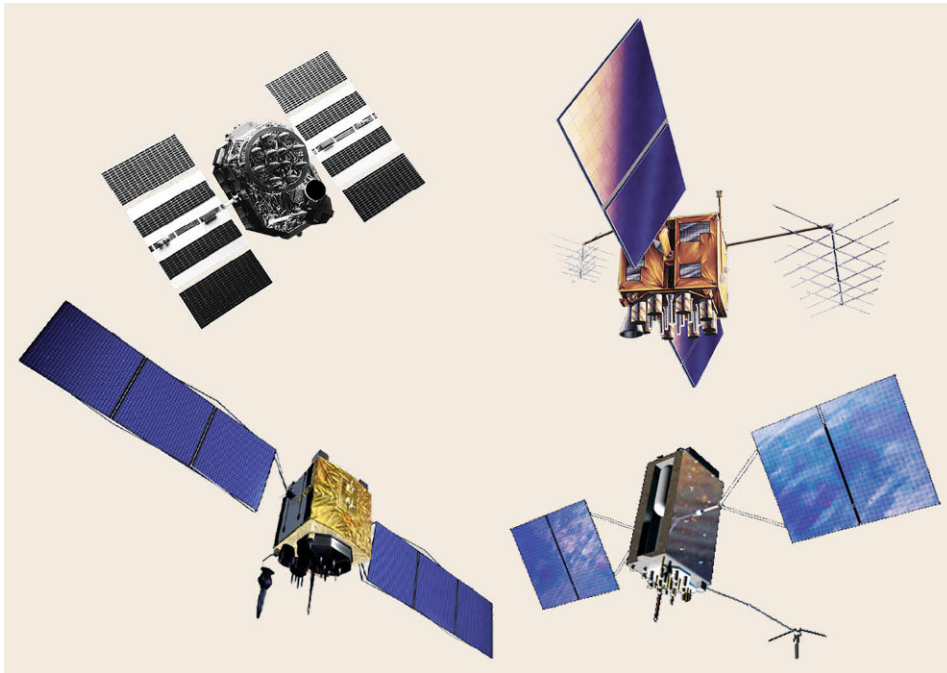


Fig. 7.3 The GPS satellite family: Block IIA (top left), Block IIR (top right), Block IIF (bottom left), GPS III (bottom right) (courtesy of USAF)

To accomplish this objective, less gain is provided at boresight (where the surface of the Earth is closer and thus path loss is less) than toward off-boresight angles of 13.8° , which corresponds to the limb of the Earth.

Ultra-high frequency (UHF) crosslinks provide a means to relay nuclear detection (NUDET) sensor data between satellites. Sensors related to the NUDET mission, and also an antenna for S-band tracking, telemetry, and control (TT and C), may be seen below the L-band Earth-coverage antenna on the bulkhead of the satellite in Fig. 7.3.

Two Block IIA satellites – SVNs 35 and 36 – carried satellite laser retroreflectors. Both of these satellites are now decommissioned.

Block IIR and IIR-M Satellites

A contract for 21 Block IIR (*R* for replenishment) GPS satellites (Fig. 7.3) was awarded to General Electric's Astro Space Division (now Lockheed Martin) in 1989 [7.11]. The last eight IIR satellites were modernized, as discussed in the subsequent section, and are now referred to as Block IIR-M [7.12]. The remaining 13 Block IIR satellites were launched from January 1997 to November 2004. The first Block IIR launch was unsuccessful and IIR-1 was destroyed on January 1997. As of today (2015), all of the remaining Block IIRs are still in operation. The Block IIR and IIR-M satellites carry three rubidium clocks. The design life is 7.5 years. Power is provided by two solar arrays with an area of 13.6 m^2 yielding 1140 W, with nickel hydrogen (NiH_2) batteries for storage.

A similar 12-element array design is used for the L-band navigation signals as was described earlier for the Block II/IIA satellites. The IIR L-band antenna gain pattern is slightly narrower than the IIA antenna gain pattern, which provides more power to terrestrial GPS users, but less power to GPS receivers on spacecraft in certain orbits. A modified antenna design was incorporated for the last four IIR and all eight IIR-M satellites that provided yet further gain increases toward the Earth (but yet further power decreases for the much smaller number of space users) [7.13].

The eight Block IIR-M satellites were launched from September 2005 through August 2009. These satellites add new civilian and military signals [7.14], which are described in detail in Sect. 7.3. The seventh Block IIR-M satellite (SVN 49) carried a demonstration payload for the third civilian GPS signal, referred to as *L5* (Sect. 7.3). Unfortunately, this demonstration payload resulted in signal reflections for the primary L-band navigation payload [7.15] and thus this satellite has been set unhealthy since its launch.

Block IIF Satellites

A contract to build the GPS Block IIF (*F* for *follow-on*) was awarded to Rockwell International (now Boeing) in 1996. The initial contract included options for up to 33 satellites, but only 12 were procured. Each Block IIF satellite [7.16, 17] (Fig. 7.3) is approximately 17.5 m from end to end. Six solar panels, populated with gallium arsenide (GaAs) solar cells with the total area of 22.2 m^2 , provide primary power. The power system, which also includes NiH_2 batteries, is capable of providing 2610 W at the end of a 12-year satellite design life. The on-orbit mass is approximately 1630 kg. The IIF satellites provide all of the navigation signals provided by the Block IIR-M satellites and additionally a new civilian signal at 1176.45 MHz (Sect. 7.3). The first Block IIF satellite was launched in May 2010. As of 2015, 9 of the 12 IIF satellites have been launched.

GPS III Satellites

In May 2008, the United States Air Force awarded a contract to Lockheed Martin to develop the third generation of GPS satellites. The contract called for the delivery of two satellites with options for up to 10 more. These satellites were originally referred to as Block III, but are now referred to as GPS III. These satellites [7.18] are anticipated to be launched beginning in 2017. Each GPS III satellite carries three rubidium clocks, and will provide a fourth civilian GPS navigation signal (Sect. 7.3), in addition to all the navigation signals broadcast by the Block IIF satellites. The on-orbit mass of the GPS III satellites is approximately 2200 kg. The power system includes four GaAs solar arrays with a total area of 28.5 m^2 and NiH_2 batteries, and is designed to provide 4480 W of power at the end of a 15 year design life. The satellite main body is approximately $3.4\text{ m} \times 2.5\text{ m} \times 1.8\text{ m}$. Later GPS III satellites are expected to reintroduce satellite laser retroreflectors, a capability not seen on GPS since the decommissioning of Block IIA satellites SVNs 35 and 36 (Sect. 7.1.2). Plans are also being made to introduce a search and rescue (SAR) payload on later GPS III satellites. This SAR payload will be interoperable with the International Cospas-Sarsat System.

Launch Operations

The Block I GPS satellites were launched from Vandenberg Air Force Base in California. All later satellites have been launched from Cape Canaveral Air Station in Florida. Atlas-E/F launch vehicles were used for the Block I satellites, and Delta II launch vehicles for the Blocks II, IIA, IIR, and IIR-M. The Atlas-F and Delta II launch vehicles (Fig. 7.4) were not powerful enough to directly place the satellites into their final circular MEO

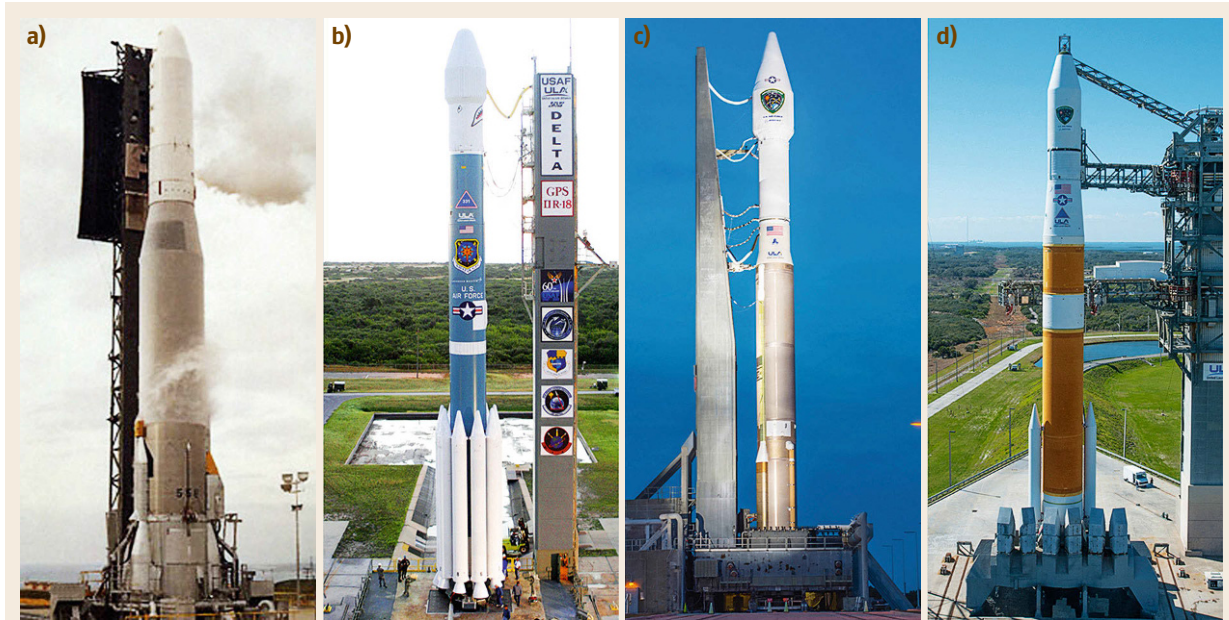


Fig. 7.4a–d The GPS launch vehicles: *Atlas–F SFSI* (a), *Delta II* (b), *Atlas V* (c), *Delta IV* (d) (courtesy of USAF (a) and United Launch Alliance (b,c,d))

orbits, but rather into highly elliptical transfer orbits with apogees near the correct final circular orbit altitude of approximately 20 000 km. Apogee kick motors (AKM) on-board the satellites were required to insert the satellites into their final circular orbits. The Block IIF satellites are being launched with evolved expendable launch vehicles (EELV) including the Delta IV and Atlas V boosters (Fig. 7.4). These launch vehicles are powerful enough to directly insert the IIF satellites into their final circular orbits without the need for AKMs. The GPS III satellites will also be launched using EELVs. However, their mass is significantly higher than that of the Block IIF satellites, and a liquid apogee engine is again required for final orbit insertion.

For launch, each GPS satellite is placed into a stowed configuration, for example, the solar arrays are folded against the side of the satellite. The satellite is mounted atop the launch vehicle within a *payload fairing*, which is a cover designed to eventually break away but to protect the satellite during ascent through the Earth's atmosphere. From this point, the launch details vary with satellite Block. As one example, the last Block IIR-M satellite launch was accomplished on August 17, 2011 using a Delta II vehicles with three stages in its 7925 configuration [7.19]. The cylindrical launch vehicle dimensions were approximately 38 m \times 2.4 m. The first (main) stage was 26 m in length at the bottom of the launch vehicle and burned a mixture of rocket propellant and liquid oxygen stored in large on-board

tanks. The first stage was assisted by nine smaller solid rocket motors attached to its base. At liftoff ($t = 0$ s), both the main stage and six of the solid rocket motors were fired. After about a minute, the six ground-start solid rocket motors were expended and jettisoned, and the remaining three solid rocket motors were ignited. These remaining solid rocket motors burnt out at $t = 128$ s and were jettisoned shortly thereafter. The main engine cut off at $t = 263$ s and then separated from the rest of the launch vehicle. The second stage, which was a restartable, hypergolic rocket engine manufactured by Aerojet, was ignited at $t = 277$ s. The second stage included the flight guidance electronics. The payload fairing was jettisoned 20 s later at $t = 297$ s. After a series of maneuvers and a first cut-off and restart, the second stage was cutoff a final time and jettisoned at approximately $t = 1$ h. The third stage, a solid rocket motor, was then ignited for less than 2 min before being cut off. The satellite was separated from the third stage at $t = 1$ h 8 min, and reached first apogee of the transfer orbit at $t = 4$ h 3 min. On August 19, the IIR-M satellite used its AKM to raise itself into its final circular orbit.

Attitude Control

As the GPS satellites orbit the Earth, they are 3-axis stabilized to simultaneously point the L-band navigation antenna toward Earth and the solar arrays toward the Sun. Earth-pointing accuracy is typically well below 0.5° . Continuous yawing around the L-band antenna

boresight direction is required to achieve these objectives. Attitude control is accomplished using momentum wheels, magnetic torquers, and Earth–Sun sensors in system designs that vary with satellite Block. Modeling satellite yaw, particularly during eclipse season

when a portion of the satellite orbit is hidden from the Sun by the Earth, is of importance for a number of high-precision positioning applications (Chaps. 19 and 25). Suitable yaw models for the Block II/IIA, IIR, and IIF satellites may be found in [7.20–22].

7.2 Control Segment

7.2.1 Overview

The GPS satellites are monitored, commanded, and controlled by a ground network referred to as the GPS CS. The CS includes a master control station (MCS) at Schriever Air Force Base (AFB) in Colorado (Fig. 7.5), and a global network of monitor stations and ground antennas (Fig. 7.6). The monitoring sta-



Fig. 7.5 Schriever air force base, Colorado, site of the GPS MCS (courtesy of USAF)

tions include high-precision GPS receivers that track the L-band navigation signals broadcast by each visible satellite using antennas with *hemispherical* (i. e., all directions above the local horizon) gain antennas. Measurements are continuously communicated back to the central MCS.

The MCS includes the computing facilities for processing these measurements to generate estimates of the GPS satellite positions, satellite velocities, clock errors, and clock drift rates. The MCS is manned 24 h a day, 7 days of week, and the trained US Air Force personnel that work here also monitor the health of the GPS satellites and manage satellite maneuvers and navigation upload data.

High-gain, highly directional ground antennas are used to read telemetry data from the GPS satellites and to provide command and navigation data uplinks. This TT&C function is accomplished using signals compliant with the Air Force Space Ground Link Subsystem (SGLS) channel plan. An uplink frequency of 1783.74 MHz and a downlink frequency of 2227.50 MHz are used. The ground antennas are large

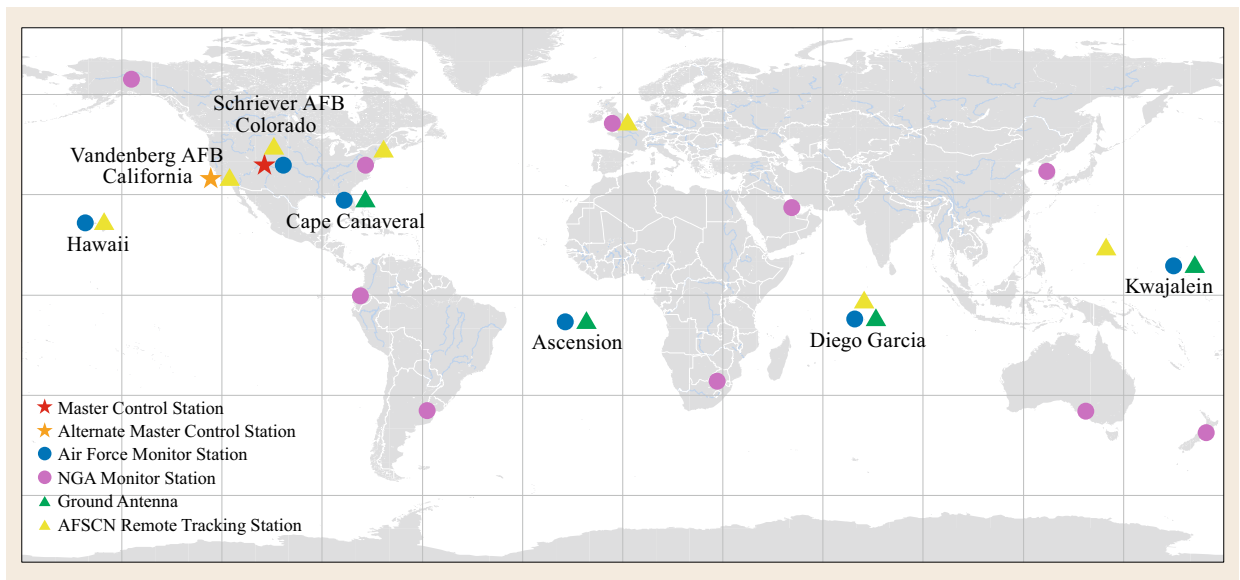


Fig. 7.6 GPS control segment

(≈ 10 m) and must be pointed to an individual GPS satellite, so scheduling is required by the operators at the MCS.

Routine navigation data uploads are performed as follows. First, as mentioned above, the MCS processing facilities continually estimate satellite position and clock parameters. These facilities also predict these quantities some number of days into the future (e.g., where the satellite is expected to be in the future, and what clock error is expected in the future). The exact prediction timeframe varies from satellite block to block. Typically once per day, the MCS schedules ground contact with each GPS satellite. Clock, ephemeris, and other data are uploaded to the satellite sufficient for that satellite to broadcast navigation data (Sect. 7.4.4) to the GPS users over one or more days without further ground contact from the CS. The operators are quite busy over the course of a day, since in recent years 31 satellites need to be uploaded so an upload must be accomplished at least once every 45 min so that the entire constellation is provided an upload in a day. Furthermore, problematic satellites with poor-performing clocks may need more frequent uploads (e.g., twice/day) to maintain desired levels of accuracy, and any anomalous satellite behavior needs to be managed.

7.2.2 Evolution of Capabilities

Since the GPS program began, the CS has undergone a number of significant transitions. The contract for the initial control segment (ICS) was awarded to General Dynamics in September 1974. The ICS included four monitor stations (Hawaii, Alaska, Guam, and Vandenberg AFB), which provided measurements to an MCS at Vandenberg AFB. A single upload station (ULS) was located at Vandenberg and provided navigation data to the GPS satellites. The ICS was operated from 1978 to 1985 primarily to support system development and user equipment testing at Yuma, Arizona [7.23].

In September 1980, a contract was awarded to IBM Federal Systems to develop the operational control segment (OCS) [7.2, 3, 24]. The OCS became operational in 1985 and was used until 2007. The OCS originally included six monitor stations (labeled as *Air Force Monitor Station* in Fig 7.6) and four dedicated ground antennas (Fig 7.6). Eight of the ten additional monitor stations, operated by the National Geospatial-Intelligence Agency (NGA), as shown in Fig 7.6, were added in 2005–2006 (all those shown in Fig 7.6 except Alaska and South Korea). The addition of the eight NGA monitor stations was part of a program referred to as the Legacy Accuracy Improvement Initiative (L-AII) that also included improvements to the OCS data pro-

cessing algorithms [7.25]. The MCS was originally situated at Vandenberg AFB, but then moved to Falcon AFB (now Schriever AFB) in 1986. Data processing at the MCS was performed by an IBM mainframe computer. A backup MCS in Gaithersburg, Maryland became operational several years later.

A contract was awarded to Lockheed Martin in 1996 to modernize the OCS. Lockheed Martin had acquired the former IBM Federal Systems (that had developed the original OCS) by that point in time. In 2000, as part of the GPS Block IIF satellite procurement, Boeing became the prime contractor for the modernized OCS, with Lockheed Martin as a subcontractor, to develop and deploy the modernized control segment that is still in use today. The modernized OCS hardware and software suite is referred to as the architecture evolution plan (AEP) [7.24]. AEP became operational in September 2007. It originally included the same monitor station, ground antenna, and MCS sites as the original OCS. The MCS uses a distributed set of Sun workstations. When it first became operational in 2007, AEP introduced an alternate master control station (AMCS) at Vandenberg AFB and an enhanced level of compatibility with the Air Force Satellite Control Network (AFSCN). Although not dedicated to GPS, AFSCN ground antennas can be utilized if needed for commanding or uploading data to GPS satellites, as well as downloading telemetry data. Two additional NGA monitor stations (Alaska and South Korea) were added to AEP in 2008. The final set of ground assets is shown in Fig. 7.6.

A further evolution of the CS is planned in a program referred to as the next generation operational control segment or OCX for short [7.26, 27]. The OCX program contract was awarded to Raytheon in 2010. OCX will be deployed in blocks. Block 0 will provide the Air Force the ability to support launch and check-out operations for the GPS III satellites. Block 1 will add functionality to permit the transition from AEP to OCX. At present, the cutover from AEP to OCX Block 1 is anticipated to take place in 2018.

7.2.3 Operations

The CS monitor stations continuously track the GPS L1 and L2 P(Y)-code signals from visible satellites using keyed, geodetic-quality receivers. Each monitor station includes additional components including a GPS antenna, redundant cesium clocks, meteorological sensors (not currently used), work stations, and communications equipment. Pseudorange, carrier-phase measurements, demodulated navigation data bits, and signal reception quality indicators are sent to the MCS from each monitor station every 1.5 s.

The MCS (or AMCS) corrects the monitor station measurements for various errors, edits the data, and then feeds the results into an extended Kalman filter. Ionospheric delays are removed using standard linear combinations (Sect. 7.4.3) of the pseudorange and carrier-phase measurements. Tropospheric delays are corrected using the Niell–Saastamoinen model [7.28]. Data editing [7.2, 3] is performed to protect the Kalman filter from suspect measurements.

The measurement update rate and the output estimate rate in the MCS Kalman filter is 15 min. The outputs of the Kalman filter, referred to as *estimated states*, include:

- For each operational GPS satellite: 3 position coordinates and 3 velocities, all in an earth centered inertial (ECI) coordinate system; 2 solar pressure parameters; 3 clock parameters (clock error and its first and second derivative)
- For each monitor station: Tropospheric wet height and 2 clock parameters (clock error and its first derivative).

For the entire operational GPS constellation (31 satellites in 2015) and monitor station network (16 in 2015), the Kalman filter is estimating a very large number of parameters every update cycle – over 380. As currently implemented in AEP, the Kalman filter partitions the states to reduce complexity. For numerical stability, the upper-diagonal (U-D) form of the Kalman filter equations is used [7.3].

Once the CS has the Kalman filter estimates of the satellite position coordinates and clock errors, additional processing is required to predict into the future how these parameters will progress with time. GPS was designed to continue to provide navigation services for extended periods even if the CS was no longer to provide satellite uploads. The Block IIA, IIR, IIR-M, and IIF satellites are uploaded with a minimum of 60 days of navigation data [7.29]. Both the Kalman filter estima-

tion of the satellite positions and the prediction forward in time of these positions require elaborate force models within the CS. At the present time, the GPS satellite force models used by the MCS include [7.3, 24, 25]:

- Earth gravitational model (EGM) of 1996, 12×12 spherical harmonic expansion coefficients
- Sun–Moon gravity
- Solid Earth tide modeling per [7.30]
- Jet Propulsion Laboratory (JPL) empirical solar radiation pressure models
- Zonal and diurnal–semidiurnal tidal corrections to Earth orientation parameters.

When OCX becomes operational, improved measurement processing and improved satellite force models are expected to significantly reduce the control segment’s contribution to the overall GPS error budget [7.26, 27].

The GPS CS includes interfaces with multiple external systems. NGA provides the data from the NGA monitor stations, as well as Earth orientation parameter predictions (EOPP) that are needed by the CS Kalman filter for ECI-ECEF coordinate transformations [7.31]. The United States Naval Observatory (USNO) provides time services that are needed to keep GPS time in synchronization with coordinated universal time (UTC). These service including maintenance of the USNO alternate master clock (AMC) that was installed and became operational at Schriever AFB in 1996. The Jet Propulsion Laboratory (JPL) provides estimates of GPS satellite hardware-induced group delay biases between signals on different carrier frequencies [7.32].

Lastly, the GPS CS generates a number of data files for use by the GPS community. These include constellation almanacs and notice advisory to NAVSTAR users (NANU). The data files are made available to the general public by the US Coast Guard, and descriptions of the files are provided in [7.33].

7.3 Navigation Signals

This section provides an overview of the GPS navigation signals, present and future. The timing for all components of all of the GPS navigation signals are coherently derived from an onboard frequency synthesizer driven by the active atomic clock. This onboard synthesizer is designed to produce a 10.23 MHz frequency as apparent to a user on or near the surface of the Earth. An observer moving along with the satellite would see the fundamental clock frequency to run at a slightly slower rate, approximately

10.229999995453 MHz, due to the combined effects of special and general relativity [7.2].

7.3.1 Legacy

The oldest 15 of the 31 operational satellites, comprising Block IIA and IIR vehicles launched up through 2004, only broadcast what are now referred to as the *legacy* GPS signals. The legacy GPS signals include the coarse/acquisition- (C/A-) code signal on the link 1 (L1)

carrier frequency of 1575.42 MHz and the precision- (P-) code signal on both L1 and the link 2 (L2) frequency of 1227.6 MHz [7.29]. The C/A-code is open (unencrypted). The P-code signal is only intended for authorized (military) use and is normally encrypted. When the P-code is in this encrypted mode of operation, it is formally referred to as the *Y-code*. In either mode of operation, the signal is most commonly referred to as the *P(Y)-code*.

Both legacy GPS signals are generated using direct sequence spread spectrum (DSSS) modulation, illustrated in Fig. 7.7. A DSSS signal may be formed as the product of three components:

1. A radio frequency (RF) carrier
2. A data waveform
3. A spreading waveform.

For C/A-code or P(Y)-code, the RF carrier is simply a pure sinusoid at L1 or L2. The data waveform is a series of contiguous 20 ms, unit amplitude rectangular pulses generated at 50 Hz with the polarity of the pulses determined by the binary, 50 bps navigation data to be conveyed from the satellite to the user. The spreading waveform is a contiguous series of rectangular pulses generated using a deterministic, digital pseudo-random noise (PRN) code. The minimum period between transitions in the spreading waveform is referred to as a *chip* denoted T_c and the reciprocal of this period as the *chipping rate*, R_c , which is the clock rate of the binary PRN code.

The PRN code for each C/A-code signal is taken from the family of length-1023 Gold codes [7.34], and is generated at 1.023 MHz. A unique PRN is used for each signal-type broadcast by each GPS satellite. The PRN codes for the Y-code are generated at 10.23 MHz using private key encryption. Both the algorithm and the keys for Y-code are only available to authorized (e.g., military) users.

It should be noted that Fig. 7.7 is not drawn to scale. There is a tremendous difference in timescales between the GPS signal components that would be difficult to illustrate on a single plot. For instance, for the C/A-code, there are 1540 cycles of the RF carrier for every single

chip in the spreading waveform, and 20 460 chips in the spreading waveform for every one data bit. If one visualizes the signal traveling through free space from the satellite to a user on or near the surface of the Earth, each period in the RF carrier stretches over approximately 19 cm, each C/A-code chip over 297 m, and each data bit over nearly 6000 km.

The legacy signals have baseband power spectra that are characteristic for any DSSS signals using rectangular chips, which is

$$S(f) = T_c \frac{\sin^2(\pi f T_c)}{(\pi f T_c)^2}, \quad (7.1)$$

with $T_c = 1/(1.023 \text{ MHz})$ for C/A-code, and $T_c = 1/(10.23 \text{ MHz})$ for P(Y)-code.

7.3.2 Modernized Signals

Figure 7.8 illustrates the evolution of the GPS signals over satellite blocks. The figure shows the normalized power spectra (on a logarithmic scale) of the GPS signals for each block versus frequency, beginning with the legacy signals broadcast by the Block I, II, IIA, and IIR satellites. The legacy signals exhibit the baseband power spectra provided in equation (7.1). Several of the modernized signals exhibit power spectra with the same spectral characteristics because they are also using DSSS modulations with C/A- or P(Y)-code chipping rates. Table 7.4 gives an overview of the GPS signals.

The eight Block IIR-M satellites launched between 2005 and 2009 introduced two new navigation signals – a new military signal on L1 and L2 referred to as the M code [7.35], and a new civil signal on L2 referred to as L2C [7.29, 36]. (As mentioned earlier, only seven of the eight are broadcasting usable signals today; the seventh Block IIR-M satellite is set unhealthy). The Block IIF satellites, launched since 2010 (9 thus far, with 12 in total planned), introduced a third civil signal on a new carrier frequency. Both the carrier and the signal are referred to as L5 [7.37, 38]. The GPS III satellites, which are anticipated to be launched beginning in 2017, will introduce one further signal – a fourth civil signal referred to as L1C on the L1 carrier.

Compared with the legacy signals, the modernized GPS signals have a number of advanced design features. For all of the modernized civilian signals, these features include dataless components, longer PRN codes, and various improvements to the navigation data encoding and content. L5 and L2C additionally employ secondary codes, and L5 and L1C use wider bandwidth modulations.

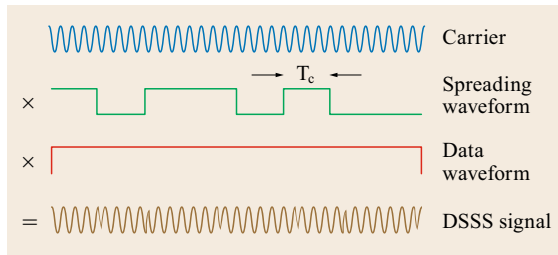


Fig. 7.7 Direct sequence spread spectrum modulation

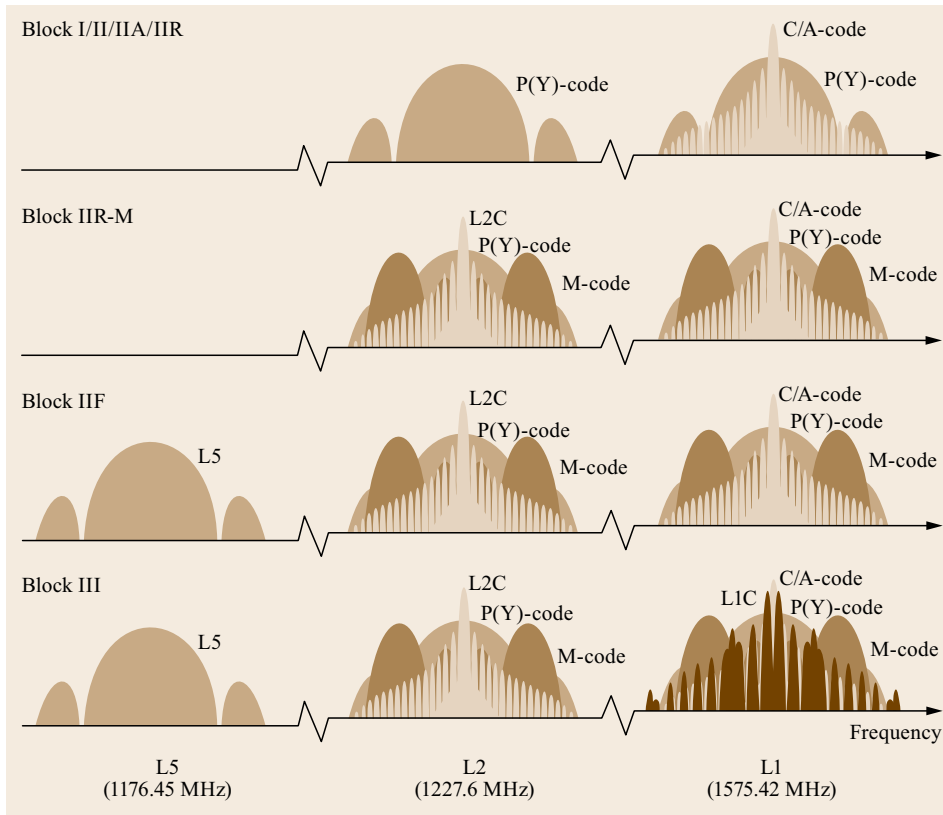


Fig. 7.8 Evolution of the GPS signals

Table 7.4 GPS signals overview. See text and Chap. 4 for a description of the various modulations forms and the meaning of the corresponding acronyms

Band	Signal	Frequency (MHz)	Code length chips	Code rate (MHz)	Data rate (bps/sps)	Modulation	Block			
							I/II/IIA/IIR	IIR-M	IIF	III
L1	P(Y)	1575.42	n/a ^a	10.23	50/50	BPSK(10)	×	×	×	×
	C/A	1575.42	1023	1.023	50/50	BPSK(1)	×	×	×	×
	L1C	1575.42	10 230	1.023	50/100	TMBOC(6,1,4/33)				×
	L1C	1575.42	10230/1800	1.023	—	TMBOC(6,1,4/33)				×
	M	1575.42	n/a ^a	5.115	n/a ^a	BOC _{sin} (10,5)		×	×	×
L2	P(Y)	1227.60	n/a ^a	10.23	50/50	BPSK(10)	×	×	×	×
	L2 CM	1227.60	10 230	0.5115	50(25)/50	BPSK(1) mux		×	×	×
	L2 CL	1227.60	767 250	0.5115	—	BPSK(1) mux		×	×	×
	M	1227.60	n/a ^a	5.115	n/a ^a	BOC _{sin} (10,5)		×	×	×
L5	I5	1176.45	10230/10	10.23	50/100	BPSK(10)			×	×
	Q5	1176.45	10230/20	10.23	—	BPSK(10)			×	×

^a Indicates nonavailability of public information for signals of regulated/military services.

A dataless component (also referred to as a *pilot*) is a portion of a GNSS signal that is not modulated by navigation data. The motivation for including a dataless component is to enable more robust tracking of the signal by a receiver in low signal-to-noise conditions. The receiver can track the RF carrier component of a pilot using a pure phase locked loop (PLL), whereas a Costas

loop is required to track a signal modulated by unknown binary data. A PLL can track a signal with approximately one-quarter the signal-to-noise ratio necessary for a Costas loop. Even though only one-half of the total power in each transmitted L2C and L5 signal is devoted to a dataless component, there is still a net 3 dB tracking robustness benefit allowing a receiver to continue

to provide measurements in the presence of more interference or more signal attenuation (due, for example, to obstructions in the line of sight between the user and satellite).

The dataless components for the modernized GPS signals are implemented differently. For L2C, a time division multiplexing scheme is used to implement the pilot component as shown in Fig. 7.9. Two unique PRN codes per satellite are generated at 511.5 kHz, which is half of the 1.023 MHz C/A-code chipping rate. The chips from the two PRN codes are alternated, that is, the first transmitted chip is from the first PRN, the second chip from the second PRN, the third chip from the first PRN, and so forth. The chips generated from the first PRN code are modulated by navigation data, and the chips generated from the second PRN code are not. The resultant signal has the same power spectrum as the C/A-code (ignoring finescale structure due to the periodic PRN codes) but only half the chips in the spreading waveform are modulated by navigation data.

To implement a pilot component for L5, two equal-power DSSS signals per satellite are broadcast in phase quadrature upon the same carrier, using unique PRN codes per signal. The inphase signal is referred to as I5 and is modulated by navigation data. The quadrature signal, known as Q5, is not modulated by the navigation data. A similar method is used to generate a pilot component for L1C, except that the two L1C components are inphase with each other and for L1C, 3/4 of the signal power is devoted to the pilot.

All the modernized GPS civil signals use PRN codes that are at least 10 times longer than those used for the legacy civilian C/A-code signal. Longer PRN codes reduce interference between signals when being processed by a receiver that is receiving signals simultaneously from multiple satellites. Whereas the

C/A-code uses length-1023 PRN codes, L5 and L1C use two length-10 230 codes for each satellite (one for the data component and one for the dataless component of each signal). L2C uses two different PRN lengths per satellite for its data and dataless components. A length-10 230 PRN, referred to as the moderate-length code (CM) is used for the data component, and a long code (CL) of length-767 250 is used for the pilot. The L2C and L5 PRN codes can be generated by linear feedback shift registers. The L1C PRN codes are constructed in a more complicated manner [7.39]. User equipment can either replicate this construction process in software/hardware or, alternatively, may simply store these length-10 230 PRNs in memory.

The L5 and L1C signal components are further modulated by *secondary codes* (also referred to as *overlay codes*). Secondary codes reduce interference between GNSS signals and also facilitate robust data bit synchronization within GNSS receivers. The specific secondary codes used for L5 are referred to as Neuman–Hofman codes after the researchers that identified them for another application over 40 years ago [7.40]. The Neuman–Hofman code for I5 is 10 bit in length, and the code for Q5 is 20 bit in length. Like all binary synchronization codes, the Neuman–Hofman codes are generated at the PRN code repetition interval (1 ms) and simply either leave each repetition of the PRN code as it originally was or inverts it, depending on whether the corresponding synchronization code value is a digital 0 or 1. For instance, the Neuman–Hofman code for I5 is 0000110101. For every 10 repetitions of the I5 PRN code, the first 4 repetitions are transmitted as is, the 5th and 6th inverted, the 7th as is, the 8th inverted, the 9th as is, and the 10th inverted.

The M code uses a variant of DSSS modulation referred to as a binary offset carrier (BOC) [7.41]. BOC

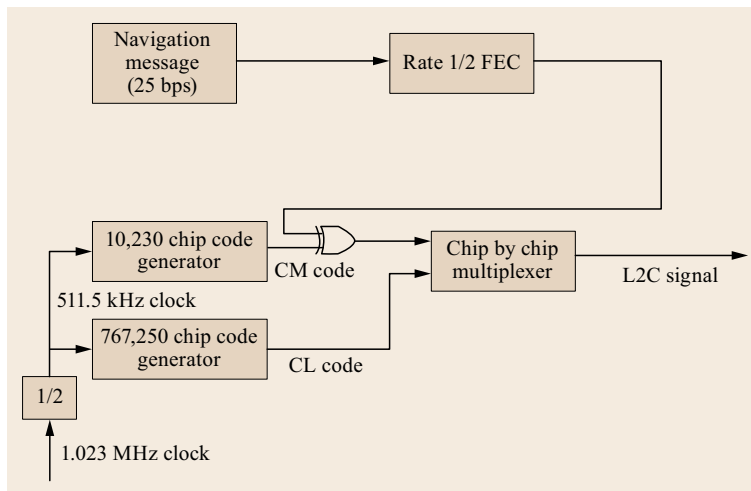


Fig. 7.9 Baseband L2C signal generation

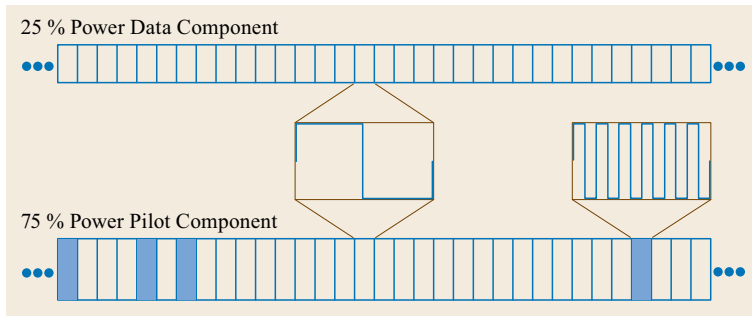


Fig. 7.10 L1C signal design

modulation adds a fourth component – a deterministic square wave – to the three normal DSSS components shown in Fig 7.1. For M-code, the spreading waveform is generated with a 5.115 MHz chipping rate and the square-wave component is clocked at 10.23 MHz so that there are two cycles of square wave per spreading waveform chip. The addition of the square-wave component gives rise to a power spectrum (Fig. 7.8) that resembles the superposition of two equal-power DSSS power spectra (7.1) at the same chipping rate (5.115 MHz), offset in frequency from the carrier by plus and minus the square-wave frequency. The effect of the square-wave component is thus similar in effect to amplitude modulation double sideband, which uses a sinusoid rather than a square wave. The notation $\text{BOC}(m, n)$ is widely used to refer to a BOC modulation with an $m \times 1.023$ MHz square-wave frequency and an $n \times 1.023$ MHz chipping rate. Using this notation, the M-code may be stated to employ a $\text{BOC}(10, 5)$ modulation.

As is well known, in estimation theory [7.42], a receiver's ability to precisely estimate the time of arrival of an arbitrary signal in the presence of additive white Gaussian noise is well predicted by the root-mean-square (RMS) signal bandwidth. The RMS signal bandwidth is defined as the integral of the signal's power spectrum weighted by the squared frequency deviation from the carrier frequency. The larger a signal's RMS bandwidth, the more precise the time-of-arrival measurement. Thus, BOC signals enable more precise pseudorange measurements than would be obtained by a DSSS signal with the same chipping rate and signal-to-noise ratio. L5 provides a signal with a large RMS bandwidth for civilian users, but does so using ordinary DSSS with a 10 times higher chip rate than the C/A-code.

The L1C signal [7.43, 44] also provides an RMS bandwidth greater than that of the C/A-code or L2C. L1C uses a mixture of two BOC modulations referred to

Table 7.5 GPS civil signal minimum specified power levels

Signal	Minimum specified received power (dBW)
C/A-code	−158.5
L2C	−160 (IIR-M/IIF), −158.5 (III)
L5	−154.9 (IIF), −154 (III)
L1C	−157

as multiplexed BOC (MBOC) [7.45]. The construction of L1C is illustrated in Fig. 7.10. The L1C data component uses $\text{BOC}(1, 1)$ modulation and has one-fourth of the overall L1C power dedicated to it. The pilot component is more complicated and is perhaps most easily understood through the viewpoint of its generation, employing DSSS but with 1.023 MHz symbols that are not rectangular pulses but rather either one or six cycles of a square wave – $\text{BOC}(1, 1)$ or $\text{BOC}(6, 1)$ symbols, respectively. As shown in Fig. 7.10, 29 of every 33 symbols in the spreading waveform are $\text{BOC}(1, 1)$ and 4 of 33 are $\text{BOC}(6, 1)$. Since three-fourth of the L1C power is in the pilot, overall $\frac{3}{4} \times \frac{4}{33} = \frac{1}{11}$ of the signal power is devoted to $\text{BOC}(6, 1)$ modulation and the remaining $\frac{10}{11}$ of the power is $\text{BOC}(1, 1)$.

7.3.3 Power Levels

The minimum specified received power levels for the GPS civil signals are summarized in Table 7.5. These levels are applicable for satellites at or above 5° elevation angle and assume that the user is located on or near the surface of the Earth with a 3 dBi linearly polarized antenna at worst normal orientation. To achieve the C/A-code received power requirement, each GPS satellite must broadcast this signal with an effective isotropic radiated power (EIRP) of around 26.4 dBW. Assuming a Block IIA antenna (Sect. 7.1.2) with 13.2 dBi gain towards the edge of Earth, this EIRP requires that approximately 20 W of C/A-code power be provided to the satellite antenna input port.

7.4 Navigation Data and Algorithms

7.4.1 Legacy Navigation (LNAV) Data Overview

The content of the 50 bps navigation data upon the C/A and P(Y)-code signals is summarized in Fig. 7.11. The data is organized into 300 bit subframes, comprising ten 30 bit words. Each 30 bit word conveys only 24 information bits. The remaining 6 bit are used for parity so that user equipment can detect transmission errors (Sect. 7.4.2).

Each subframe begins with a telemetry word (TLM) and handover word (HOW) with substantive content

summarized in Table 7.6. Subframe 1 (Sect. 7.4.3) includes clock correction data to relate the time kept by the satellite clock to the common GPS system timescale. Subframe 1 also includes information on the broadcasting satellite’s health and predicted accuracy. Subframes 2 and 3 provide ephemeris data, to be used for determining the satellite’s precise position. Subframes 4 and 5 provide less important data at a slow rate. These subframes are cycled through 25 pages each and include almanac and health data for the remaining GPS satellites as well as ionospheric correction data and the time difference between UTC and GPS system time.

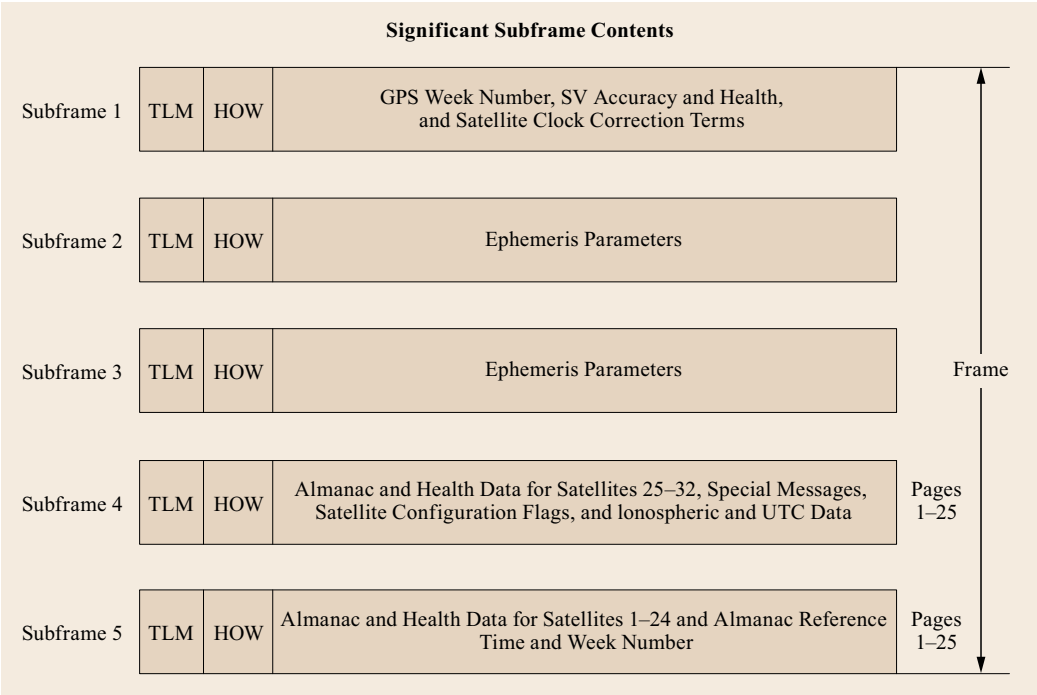


Fig. 7.11 Legacy GPS signal navigation data

Table 7.6 Substantive TLM and HOW word content

Word	Subframe bit number	Parameter	Description
TLM	1–8	Preamble	Fixed bit pattern 10001011 to help user equipment synchronize with navigation data subframes
TLM	9–22	TLM message	Reserved for authorized users and control segment
TLM	23	Integrity status flag	Today set to 0. In future, may be set to 1 to indicate a higher level of integrity [7.29, Sect. 20.3.3.1]
HOW	1–17	Time of week count	Provides a timestamp (the integer number of 1.5 s epochs that have elapsed in the GPS timescale since Saturday night at midnight) corresponding to the leading edge of the next subframe
HOW	18	Alert flag	When set, use this satellite at your own risk
HOW	19	Anti-spoof flag	When set, indicates that the P-code is encrypted into the Y-code
HOW	20–22	Subframe identification (ID)	Indicates which of the five subframes is being broadcast following the HOW

The first broadcast of the navigation data includes the Subframes 1–3 data followed by the first page of Subframe 4 and the first page of Subframe 5. Subframes 1–3 are repeated, followed by the second page of Subframe 4 and the second page of Subframe 5, and so on. It takes 12.5 min for the entire set of data to be broadcast, but during this period typically the same critical clock and ephemeris data will be received 25 times within Subframes 1–3.

7.4.2 LNAV Error Detection Encoding

An extended Hamming block code is utilized to generate each 30 bit word. This code is referred to as a (32, 26) code because it is generated using 26 information (source) bits (the 24 information bits to be conveyed in the word, plus the last 2 bits of the previous 30 bit word) and results in a 32 bit output. The 32 bit are truncated to 30 bit by discarding 2 bit. The overall algorithm to produce the transmitted 30 bit word is shown in Fig. 7.12. Note that the encoding is *systematic*, that is, the information bits are included in the transmitted data stream, with one twist – if the last bit in the preceding word is a 1, then all 24 information bits are inverted. For each 30 bit word, the parity scheme enables the user equipment to detect all possible combinations of up to 3 bit errors.

One method that the user equipment can use to detect parity is as follows. First, read the first 24 bit of the word as transmitted. If the 30th bit of the previous

word is 1, invert these 24 bit. Apply the parity encoding equations as shown in Fig. 7.12 and compare with the transmitted parity. If there are any differences, a transmission error has occurred, so disregard the data.

7.4.3 LNAV Data Content and Related Algorithms

The following subsections provide an overview of the data content of the 50 bps LNAV data conveyed by the GPS C/A-code and P(Y)-code signals, as well as the associated algorithms. For details, the reader is referred to [7.33].

Subframe 1

Subframe 1 provides the data parameters listed and described in Table 7.7. The last four parameters are used by receivers to correct for the offset between the broadcasting satellite clock (referred to as *SV time*) and the GPS time scale. GPS receivers form raw pseudorange measurements by differencing the transmit time of each satellite signal (determined by the phase of the received PRN code), t_{SV} , from reception time (determined by the receiver clock), t_r

$$\rho = c(t_r - t_{SV}), \quad (7.2)$$

where c is the speed of light in a vacuum (299 792 458 m/s). Dual-frequency GPS equipment will measure raw pseudoranges on both L1 and L2 from

$$\begin{aligned} D_1 &= d_1 \oplus D_{30}^* \\ D_2 &= d_2 \oplus D_{30}^* \\ D_3 &= d_3 \oplus D_{30}^* \\ &\vdots \\ D_{24} &= d_{24} \oplus D_{30}^* \\ D_{25} &= D_{29}^* \oplus d_1 \oplus d_2 \oplus d_3 \oplus d_4 \oplus d_5 \oplus d_6 \oplus d_{10} \oplus d_{11} \oplus d_{12} \oplus d_{13} \oplus d_{14} \oplus d_{17} \oplus d_{18} \oplus d_{20} \oplus d_{23} \\ D_{26} &= D_{30}^* \oplus d_2 \oplus d_3 \oplus d_4 \oplus d_6 \oplus d_7 \oplus d_{11} \oplus d_{12} \oplus d_{13} \oplus d_{14} \oplus d_{15} \oplus d_{18} \oplus d_{19} \oplus d_{21} \oplus d_{24} \\ D_{27} &= D_{29}^* \oplus d_1 \oplus d_3 \oplus d_4 \oplus d_5 \oplus d_7 \oplus d_8 \oplus d_{12} \oplus d_{13} \oplus d_{14} \oplus d_{15} \oplus d_{16} \oplus d_{19} \oplus d_{20} \oplus d_{22} \\ D_{28} &= D_{30}^* \oplus d_2 \oplus d_4 \oplus d_5 \oplus d_6 \oplus d_8 \oplus d_9 \oplus d_{13} \oplus d_{14} \oplus d_{15} \oplus d_{16} \oplus d_{17} \oplus d_{20} \oplus d_{21} \oplus d_{23} \\ D_{29} &= D_{30}^* \oplus d_1 \oplus d_3 \oplus d_5 \oplus d_6 \oplus d_7 \oplus d_9 \oplus d_{10} \oplus d_{14} \oplus d_{15} \oplus d_{16} \oplus d_{17} \oplus d_{18} \oplus d_{21} \oplus d_{22} \oplus d_{24} \\ D_{30} &= D_{29}^* \oplus d_3 \oplus d_5 \oplus d_6 \oplus d_8 \oplus d_9 \oplus d_{10} \oplus d_{11} \oplus d_{13} \oplus d_{15} \oplus d_{19} \oplus d_{22} \oplus d_{23} \oplus d_{24} \oplus \end{aligned}$$

Where

d_1, d_2, \dots, d_{24} are the source data bits;
the symbol \star is used to identify the last 2 bits of the previous word of the subframe;
 $D_{25}, D_{26}, \dots, D_{30}$ are the computed parity bits;
 $D_1, D_2, \dots, D_{29}, D_{30}$ are the bits transmitted by the SV;
 \oplus is the “modulo-2” or “exclusive-or” operation.

Fig. 7.12 LNAV data parity encoding (after [7.33])

each satellite and combine these in a manner that removes the group delay effects of the ionosphere

$$\rho = \frac{\rho_{L2} - \gamma \rho_{L1}}{1 - \gamma}, \quad (7.3)$$

where $\gamma = (f_{L1}/f_{L2})^2$, f_{L1} is the L1 carrier frequency (1575.42 MHz), and f_{L2} is the L2 carrier frequency (1227.6 MHz). This linear combination is referred to as the *ionospheric-free* pseudorange. A pseudorange correction, $\Delta\rho$, that can be very large (up to 300 km) is added to the ionospheric-free pseudorange using the last four parameters listed in Table 7.7

$$\Delta\rho = c [a_{f0} + a_{f1}(t_{SV} - t_{oc}) + a_{f2}(t_{SV} - t_{oc})^2]. \quad (7.4)$$

Single-frequency equipment must apply an additional correction to account for the fact that there are generally group delay biases in the satellite transmission chain between the L1 and L2 signals, and additionally the CS is determining the clock corrections using ionospheric-free pseudorange measurements. L1-only users apply an additional correction (to be subtracted from the pseudorange)

$$\Delta\rho = c T_{GD} \quad (7.5)$$

and L2-only users apply the additional correction (again to be subtracted from the pseudorange)

$$\Delta\rho = c \gamma T_{GD}. \quad (7.6)$$

The accuracy of the broadcast clock corrections, averaged across all operational GPS satellites for the time period from 2008 to 2014 is assessed in [7.46] to be at the 50 cm level (68%) and 1.85 m (95%).

If the satellite clock deviates from GPS time by more than can be corrected by the clock correction parameters (approximately ± 1 ms), the satellite may be set unhealthy by the CS to slew the satellite clock back into range [7.47]. Such maintenance action is required typically about once per year for the Block IIA satellites. The clocks on later-generation satellites (Block IIR and beyond) can be controlled in phase, frequency, and frequency drift. For these satellites, after initialization, the CS typically makes adjustments in the frequency drift only while the satellite is still set healthy which largely obviates the need for this type of maintenance action.

Subframes 2–3

Subframes 2–3 are used to convey ephemeris data to the user. The ephemeris parameters are listed in Table 7.8, along with their number of bits, scale factor (magnitude of the least significant bit (LSB)), and

range. The ephemeris parameters include six traditional Keplerian elements (Chap. 3) with associated reference time, t_{oc} :

- Semimajor axis, A – provided as the square-root of A
- Eccentricity, e
- Mean anomaly, M_0 – valid for the reference time, t_{oc}
- Longitude of ascending node, Ω_0 – valid for the GPS weekly epoch (time of week = 0 s)
- Inclination angle, i_0 – valid for the reference time, t_{oc}
- Argument of perigee, ω .

All but one of the remaining parameters in Table 7.8 (the exception is IODE, which is described below) provide corrections to the elliptical, stationary orbits that are provided by the Keplerian elements. These are necessary, since as discussed in Chap. 3, Keplerian (elliptical) orbits are only perfectly valid for the two-body problem and in reality there are many more forces acting upon the GPS satellites. The supplied corrections include:

- Cosine and sine harmonic corrections to the argument of latitude, orbital radius, and inclination angle. The amplitudes of these corrections are provided by C_{us} , C_{uc} (sine and cosine argument of latitude, respectively); C_{rs} , C_{rc} (sine and cosine orbital radius, respectively); C_{is} , C_{ic} (sine and cosine inclination angle, respectively).
- Mean motion difference from computed value, Δn .
- Rate of right ascension, $\dot{\Omega}$
- Rate of inclination angle, \dot{I} .

The accuracy of the broadcast ephemeris data, averaged across all operational GPS satellites for the time period from 2008 to 2014 is assessed in [7.46] to be 18, 98, and 60 cm (68%, for radial, along-track, and cross-track directions, respectively) and 0.43, 2.23, 1.25 m (95%, for radial, along-track, and cross-track directions, respectively).

The final Subframes 2–3 parameter is issue-of-data ephemeris (IODC). IODE is an 8 bit unsigned integer broadcast in both Subframes 2 and 3 that is intended to provide a means for user equipment to detect when changes have been made to the broadcast ephemeris data. IODE is set equal to the 8 LSBs of the 10 bit IODC (Sect. 7.4.3) for the same data set. By specification [7.29], the transmitted IODE will be different from any value transmitted by the SV over the preceding 6 h.

The broadcast ephemeris parameters are used to compute satellite (x , y , z) coordinates in the World Geodetic System of 1984 (WGS 84) coordinate system (Sect. 7.5) using the algorithm provided in Table 7.9. Importantly, as described in Sect. 7.2, the broadcast ephemeris data is generated by the GPS CS using

Table 7.7 Subframe 1 parameters

Subframe bit number	Parameter	Description
61–70	Week number	10 bit integer (0–1023) that counts weeks elapsed since midnight January 5, 1980 (in the GPS timescale). A rollover occurred at midnight August 21, 1999, and the next rollover will occur at midnight April 6, 2019
71–72	Code on L2	2 bit to indicate whether PRN on L2 is P(Y) (01) or C/A (10). This reflects a legacy satellite capability to transmit either the P(Y)- or C/A-code on L2. To date, P(Y) has been the normal mode of operation for all GPS satellites. The remaining 2 bit combinations are reserved
73–76	User range accuracy (URA) index	URA is defined to be a conservative estimate of root-mean-square user range errors (URE) due to signal-in-space error sources (i.e., satellite clock, ephemeris, and hardware group delay errors). In recent years, the 4 bit broadcast URA index for healthy satellites has typically been 0, 1, or 2 most of the time, corresponding to URA ranges of 0–2.4, 2.4–3.4, or 3.4–4.85 m, respectively. Higher URA index values are mapped to larger URA values. A URA index of 15 means that URA is unbounded and the user should use the satellite at their own risk
77–82	SV health	These 6 bit convey health information for the broadcasting SV. Currently, only two SV health bit patterns are broadcast the vast majority of the time: 000000 for healthy satellites or 111111 for satellites that are deemed unhealthy
83–84, 211–218	Issue-of-data clock (IODC)	The 10 bit IODC is an unsigned integer that changes whenever the broadcast clock parameters changes. The transmitted IODC is specified to be different from any value transmitted by the SV during the preceding 7 days
91	L2P data flag	When set, this bit indicates that the L2 P(Y)-code signal is NOT modulated by navigation data
197–204	T_{GD}	L1/L2 correction term
219–234	t_{oc}	16 bit unsigned integer multiplied by 16 s yields the SV clock correction reference time, t_{oc}
241–248	a_{f2}	8 bit signed (two's complement) integer multiplied by 2^{-55} s/s ² yields the SV clock drift rate correction, a_{f2}
249–264	a_{f1}	16 bit signed (two's complement) integer multiplied by 2^{-43} s/s yields the SV clock drift correction, a_{f1}
271–292	a_{f0}	22 bit signed (two's complement) integer multiplied by 2^{-31} s yields the SV clock bias correction, a_{f0}

Table 7.8 Subframe 2–3 parameters

Parameter	Number of bits	Scale factor (LSB)	Range	Units
IODE	8	1	0–255	Dimensionless
C_{rs}	16 ^a	2^{-5}	± 1024	m
Δn	16 ^a	2^{-43}	$\pm 3.73 \cdot 10^{-9}$	Semicircles/s
M_0	32 ^a	2^{-31}	± 1.0	Semicircles
C_{uc}	16 ^a	2^{-29}	$\pm 6.10 \cdot 10^{-5}$	rad
e	32	2^{-33}	0–0.03	Dimensionless
C_{us}	16 ^a	2^{-29}	$\pm 6.10 \cdot 10^{-5}$	rad
\sqrt{A}	32	2^{-19}	0–4096	\sqrt{m}
t_{oe}	16	2^4	0–604 784	s
C_{ic}	16 ^a	2^{-29}	$\pm 6.10 \cdot 10^{-5}$	rad
Ω_0	32 ^a	2^{-31}	± 1.0	Semicircles
C_{is}	16 ^a	2^{-29}	$\pm 6.10 \cdot 10^{-5}$	rad
i_0	32 ^a	2^{-31}	± 1.0	Semicircles
C_{rc}	16 ^a	2^{-5}	± 1024	m
ω	32 ^a	2^{-31}	± 1.0	Semicircles
$\dot{\Omega}$	24 ^a	2^{-43}	$\pm 9.54 \cdot 10^{-7}$	Semicircles/s
IDOT	14 ^a	2^{-43}	$\pm 9.31 \cdot 10^{-10}$	Semicircles/s

^a Signed (two's-complement) integer. All other values in table are unsigned integers.

a curve-fit procedure. Thus, the user equipment should use only the algorithm prescribed within [7.29] including the exact values of the constants involved:

- WGS 84 value of the Earth's gravitational constant, $\mu = 3.986005 \cdot 10^{14} \text{ m}^3/\text{s}^2$
- WGS 84 value of the Earth's rotation rate, $\dot{\Omega}_e = 7.2921151467 \cdot 10^{-5} \text{ rad/s}$
- $\pi = 3.1415926535898$.

For positioning with GPS, it is necessary to determine where the satellites were at different times in the past, that is, the locations of the satellites when they transmitted the signals that are received at a common time of reception. The transmit time, t , in Table 7.9 is typically 60–90 ms earlier than the time of reception (for a user on or near the surface of the Earth) depending on each satellite's elevation angle. When computing the time from reference epoch (third equation in Table 7.9), it is necessary to account for the weekly rollovers in GPS time of week. As one final note on Table 7.9, the sixth line requires solving Kepler's equation for eccentric anomaly. Methods for solving this equation are described in Chap. 3.

As noted in Sect. 7.3, the GPS satellite clock is intentionally set to run slow as apparent to an observer moving along with the satellite to compensate for the

effects of both special and general relativity. This compensation method is only sufficient for GPS satellites that are in their ideal, circular orbits. In practice, the GPS orbits are slightly elliptical. At times the GPS satellites are further from the Earth than nominal, in which case they are traveling slower than nominal and higher in the Earth's gravitational field than nominal. Other times, they are closer to the Earth than nominal, traveling faster than nominal and lower in the Earth's gravitational field. These deviations from the nominal orbit result in the need for changes to the special and general relativistic corrections. A pseudorange correction, $\Delta\rho$, is added to the measured pseudorange (after correction for SV clock and group delay bias as described in Sect 7.4.3) to account for this effect

$$\begin{aligned}\Delta\rho &= c \Delta t_r \\ &= c e F \sqrt{A} \sin E_k,\end{aligned}\quad (7.7)$$

$$\begin{aligned}\text{where } F &= -\frac{2\sqrt{\mu}}{c^2} \\ &= -4.442807633 \cdot 10^{-10} \frac{\text{s}}{\sqrt{\text{m}}}.\end{aligned}\quad (7.8)$$

This relativistic correction can be as large as 14 m in magnitude, assuming an orbital ellipticity of up to 0.02.

Table 7.9 Determination of GPS satellite position using broadcast ephemeris data (after [7.29])

$A = (\sqrt{A})^2$	Semimajor axis
$n_0 = \sqrt{\frac{\mu}{A^3}}$	Computed mean motion (rad/s)
$t_k = t - t_{\text{oe}}$	Time from ephemeris reference epoch
$n = n_0 + \Delta n$	Corrected mean motion
$M_k = M_0 + n t_k$	Mean anomaly
$M_k = E_k - e \sin E_k$	Kepler's equation for eccentric anomaly (rad)
$v_k = \tan^{-1} \left\{ \frac{\sin v_k}{\cos v_k} \right\} = \tan^{-1} \frac{\sqrt{1-e^2} \sin E_k / (1-e \cos E_k)}{(\cos E_k - e) / (1-e \cos E_k)}$	True anomaly
$E_k = \cos^{-1} \left\{ \frac{e + \cos v_k}{1 + e \cos v_k} \right\}$	Eccentric anomaly
$\Phi_k = v_k + \omega$	Argument of latitude
$\delta u_k = c_{\text{us}} \sin 2\Phi_k + c_{\text{uc}} \cos 2\Phi_k$	Argument of latitude correction
$\delta r_k = c_{\text{rs}} \sin 2\Phi_k + c_{\text{rc}} \cos 2\Phi_k$	Radius correction
$\delta i_k = c_{\text{is}} \sin 2\Phi_k + c_{\text{ic}} \cos 2\Phi_k$	Inclination correction
$u_k = \Phi_k + \delta u_k$	Corrected argument of latitude
$r_k = A(1 - e \cos E_k) + \delta r_k$	Corrected radius
$i_k = i_0 + \delta i_k + (\text{IDOT})t_k$	Corrected inclination
$\begin{cases} x'_k = r_k \cos u_k \\ y'_k = r_k \sin u_k \end{cases}$	Positions within orbit coordinate system
$\Omega_k = \Omega_0 + (\dot{\Omega} - \dot{\Omega}_e)t_k - \dot{\Omega}_e t_{\text{oe}}$	Corrected longitude of ascending node
$\begin{cases} x_k = x'_k \cos \Omega_k - y'_k \sin \Omega_k \\ y_k = x'_k \sin \Omega_k + y'_k \cos \Omega_k \\ z_k = y'_k \sin i_k \end{cases}$	Position coordinates in the World Geodetic System (WGS) 84

Subframes 4–5

As described in Sect. 7.4.1, both Subframes 4 and 5 are paged, with 25 pages of data each. A summary of the content of Subframes 4 and 5 is provided in Table 7.10. Many of the pages are reserved for authorized users or internal system use.

Pages 1–24 of Subframe 5 and Pages 2, 3, 4, 5, 6, 8, 9, and 10 of Subframe 4 provide *almanac* data. This data is essentially truncated and simplified (i. e., using fewer parameters) ephemeris data for all of the satellites in the constellation. The same algorithm as used for the ephemeris data (Table 7.9) can be used by the user equipment to determine the coarse positions of other satellites to aid acquisition.

Page 13 of Subframe 4 is specified to contain elements of a navigation message correction table (NMCT). The NMCT is only intended for authorized users and may be encrypted [7.33]. This table, which is also referred to as the wide area GPS enhancement (WAGE), essentially contains pseudorange corrections for all of the other GPS satellites [7.48].

For the benefit of single-frequency L1-only GPS users, Page 18 of Subframe 4 includes ionospheric delay correction parameters. These parameters provide to the single-frequency user an estimate of the vertical ionospheric delay errors in the form of a half-cosine during local daylight hours. The broadcast parameters are determined by a table lookup by the GPS MCS using daily measurements of solar flux at 10.7 cm and the day of the year [7.49].

Subframe 4 also provides data necessary to relate the GPS timescale to UTC as maintained by the US Naval Observatory – UTC (USNO).

7.4.4 Civil Navigation (CNAV) and Civil Navigation-2 (CNAV-2) Data

All of the modernized civil GPS signals provide similar navigation data content to that provided by the C/A-code and P(Y)-code signals. L1C and L5 convey

navigation data at 50 bps, and L2C at 25 bps. The navigation data format for L2C and L5 is referred to as CNAV, and the format for L1C is referred to as CNAV-2.

Several improvements relative to the C/A-code are notable. First, all the modernized signals use forward error correction (FEC) encoding such as rate 1/2 convolutional encoding (e.g., L2C and L5) or low-density parity check codes (L1C). FEC techniques take the raw navigation data bits to be conveyed to the end-user and judiciously add redundancy to them. Although the rate of binary data is increased (by a factor of 2 for rate 1/2 encoding), the receiver is able to decode the raw navigation data bits with a smaller probability of error. For the reader without a communication theory background, it may be helpful toward understanding GNSS signal specifications to know that the rate of raw navigation data, R_b , is normally specified in units of bits/second (bps) and the resultant higher data rate after FEC is applied, R_s , in units of symbols/second (sps). The net benefit of FEC is that the receiver can read the navigation data bits in poorer signal-to-noise conditions.

All of the modernized GPS civil signals also utilize an improved error detection scheme. The Hamming (32, 26) code used for the LNAV data for error detection (Sect. 7.4.2) is very inefficient in that 6 parity bits are used for every 30 bit word. The modernized civil signals use a 24 bit cyclic redundancy check (CRC) code along with a much longer word length (e.g., 300 bit for L5). This CRC scheme is much more efficient and allows receivers of the modernized signals to more robustly detect the occurrence of one or more bit errors.

The clock and ephemeris navigation data fields, although similar to those conveyed via C/A-code and P(Y)-code, have more precise least significant bits and some additional terms, for example, in the satellite ephemeris representation, to ensure that the data field design will not significantly limit pseudorange measurement precision in the future. New navigation data fields are also added to CNAV and CNAV-2 as

Table 7.10 Content of Subframes 4 and 5 (after [7.29])

Subframe	Page(s)	Data
4	1, 6, 11, 16, and 21	Reserved
	2, 3, 4, 5, 7, 8, 9, and 10	Almanac data for SV 25 through 32, respectively
	12, 19, 20, 22, 23, and 24	Reserved
	13	Navigation message correction table (NMCT)
	14 and 15	Reserved for system use
	17	Special messages
	18	Ionospheric and UTC data
	25	A-S flags/SV configurations for 32 SVs, plus SV health for SV 25 through 32
5	1 through 24	Almanac data for SV 1 through 24
	25	SV health data for SV 1 through 24, the almanac reference time, the almanac reference week number

compared to LNAV. These include intersignal corrections, Earth orientation parameters, and an improved set of differential corrections that break apart clock and

ephemeris errors as compared to the NMCT in LNAV that provides only a lumped pseudorange correction. For details, refer to [7.32, 37, 43].

7.5 Time System and Geodesy

GPS uses its own timescale, referred to as *GPS time*. GPS time was set equal to UTC as maintained by the US Naval Observatory (UTC [USNO]) on midnight on the night of January 5, 1980/morning of January 6, 1980. However, unlike UTC, GPS time is a continuous timescale, that is, there are no leap seconds. GPS time used to be maintained by a single physical clock in the GPS CS until 1985. Since the CS transitioned to the IBM-delivered system in 1985, GPS time has been created using a composite clock [7.50], that is, the weighted average of the aggregate set of atomic clocks within GPS. As discussed in Sect. 7.1, there are rubidium or cesium clocks on-board all of the GPS satellites, and as discussed in Sect. 7.2, there are additionally atomic clocks situated at the monitor stations and at the MCS.

The GPS CS steers GPS time to keep it within 1 μ s of UTC (USNO) by specification [7.29]. As noted in Sect. 7.3, the broadcast GPS navigation data includes parameters to relate GPS time and UTC (USNO). This data is specified to be accurate to within 90 ns, one-sigma [7.29]. Actual performance is much bet-

ter. GPS time is usually maintained to within 20 ns of UTC (USNO) and application of the broadcast GPS-UTC corrections typically yields UTC (USNO) within 5 ns [7.51].

As discussed in Sect. 7.4.3, each satellite has its own timescale referred to as *SV time* [7.29]. The clock correction data included within the navigation data (e.g., Subframe 1 for the legacy signals) is intended to allow the user to relate each satellite's SV time to the common GPS time. SV time may deviate from GPS time by up to 1 ms, which is the range of the Subframe 1 clock correction data.

The GPS satellite positions are provided to the user, via computations enabled by the Subframe 2 ephemeris data, within the World Geodetic System 1984 (WGS-84) reference frame. The coordinates of the GPS CS monitor stations are periodically readjusted to keep WGS-84 aligned with the International Terrestrial Reference Frame (ITRF). WGS-84 and ITRF are generally considered as being coincident within several centimeters.

7.6 Services and Performance

GPS presently provides two services – one for civilian users referred to as the standard positioning service (SPS) [7.5] and one available only to authorized users (primarily the US military, and the militaries of US allies) referred to as the precise positioning service (PPS). The United States has pledged to make the GPS SPS available for civil aviation use on a continuous worldwide basis, free of direct user fees, with a minimum of 6 years advance notice to be provided in the event that this service will be terminated. This commitment was initially made by the Administrator of the Federal Aviation Administration (FAA) in 1994 [7.52]. The commitment to provide GPS SPS service was reiterated in 2007 [7.53], with an additional commitment made at that time, to provide GPS satellite-based augmentation system (SBAS) services in North America, free of direct user charges, through the FAA's Wide Area Augmentation System (WAAS) (Chap. 12 for a description of SBASs, including WAAS).

At one time, the accuracy of the SPS was intentionally degraded using a technique referred to as selective

availability (SA), which was observed to be implemented as a pseudorandom dithering of the satellite clock that could be removed only by PPS receivers with the knowledge of the generation algorithm and cryptographic keys [7.2]. On May 12, 2000, the intentional degradation of SPS performance due to SA was ceased [7.54] and more recently, in September 2007 the United States announced that the capability to implement SA will be removed from future GPS satellite procurements [7.55].

The specified accuracy of the GPS SPS is 13 m, 95%, for horizontal positioning and 22 m, 95%, for vertical positioning [7.5]. This specification is for the signal-in-space (SIS) only (i.e., it does not include errors due to the atmosphere, multipath, or user equipment), and is based upon a global average. Actual performance is typically significantly better than the specification. For instance, the observed 95% horizontal and vertical positioning accuracies for 28 GPS SPS receivers distributed throughout North America from April 01 to June 30, 2013 were 3.0 and 4.3 m, re-

spectively [7.53]. Further, the data reported in [7.53] includes all real-world errors, whereas the accuracy specification in the SPS performance standard [7.5] only includes SIS errors. PPS users typically enjoy even better accuracies, on the order of 1–2 m, 95%, for both horizontal and vertical positioning. GPS users employing differential techniques, as described in later chapters of this text, can achieve accuracies better than 1 cm.

Although not an intended service for GPS, many civilian users are tracking the encrypted P(Y)-code signals on L1 and L2 without requiring access to

the cryptographic keys using codeless or semicodeless techniques [7.56]. A public notice was made in 2008 [7.57] to commit the US government to supporting codeless–semicodeless access to the GPS P(Y) signals until December 31, 2020. In 2015, this commitment was superseded by a new one to continue support to codeless–semicodeless receivers ... *until at least 2 years after there are 24 operational satellites broadcasting L5* [7.58]. Users of codeless–semicodeless GPS receivers are advised to transition to new user equipment utilizing the modernized GPS signals (L2C, L5).

References

- 7.1 B.W. Parkinson, S.W. Gilbert: NAVSTAR: Global positioning system – Ten years later, Proc. IEEE **71**(10), 1177–1186 (1983)
- 7.2 B.W. Parkinson, J.J. Spilker Jr.: *Global Positioning System: Theory and Applications*, Vol. I (American Institute of Aeronautics and Astronautics, Washington 1996)
- 7.3 E.D. Kaplan, C.J. Hegarty: *Understanding GPS – Principles and Applications*, 2nd edn. (Artech House, Boston/London 2006)
- 7.4 P. Misra, P. Enge: *Global Positioning System – Signals, Measurements and Performance*, Vol. 2 (Ganga Jamuna, Lincoln 2011)
- 7.5 Global Positioning System Standard Positioning Service Performance Standard, 4th edn. (US Department of Defense, Washington 2008)
- 7.6 G.B. Green, P.D. Massatt, N.W. Rhodus: The GPS 21 primary satellite constellation, *Navigation* **36**, 9–24 (1989)
- 7.7 A.B. Jenkin, R.A. Gick: Collision risk posed to the global positioning system by disposal orbit instability, *J. Spacecr. Rocket.* **39**(4), 532–539 (2002)
- 7.8 D.M. Galvin: History of the GPS space segment from block I to the new millennium, Proc. ION GPS 1999, Nashville (ION, Virginia 1999) pp. 1843–1854
- 7.9 L.A. Mallette, P. Rochat, J. White: Historical review of atomic frequency standards used in space systems – 10 year update, Proc. 38th Annu. PTI Meet., Washington DC (2006)
- 7.10 F.M. Czopek, S. Shollenberger: Description and performance of the GPS block I and II L-band antenna and link budget, Proc. ION GPS 1993, Salt Lake City, UT (ION, 1993) pp. 37–43
- 7.11 K. Kiser, S.H. Vaughan: GPS IIR joins the GPS constellation, Proc. ION GPS, Nashville, TN (ION, Virginia 1998) pp. 1915–1923
- 7.12 T. Hartman, L.R. Boyd, D. Koster, J.A. Rajan, C.J. Harvey: Modernizing the GPS block IIR spacecraft, Proc. ION GPS, Salt Lake City (ION, Virginia 2000) pp. 2115–2121
- 7.13 W. Marquis, D. Reigh: On-orbit performance of the improved GPS block IIR antenna panel, Proc. ION GNSS, Long Beach (ION, Virginia 2005) pp. 2418–2426
- 7.14 J.A. Rajan, J.A. Tracy: GPS IIR-M: Modernizing the signal-in-space, Proc. ION NTM, Anaheim (2003) pp. 484–493
- 7.15 S. Ericson, K. Shallberg, C. Edgar: Characterization and simulation of SVN49 (PRN01) elevation dependent measurement biases, Proc. ION ITM, San Diego (ION, Virginia 2010) pp. 963–974
- 7.16 S.C. Fisher, K. Ghassemi: GPS IIF – The next generation, Proc. IEEE **87**(1), 24–47 (1999)
- 7.17 M. Braschak, H. Brown Jr., J. Carberry, T. Grant, G. Hatten, R. Patocka, E. Watts: GPS IIF satellite overview, Proc. ION GNSS, Portland (ION, Virginia 2010) pp. 753–770
- 7.18 W. Marquis, S. Michael: GPS III – Bringing new capabilities to the global community, *Inside GNSS* **6**(5), 34–48 (2011)
- 7.19 GPS IIR-21 (M), Mission Book (United Launch Alliance, Littleton, Colorado 2009) <http://www.ulalaunch.com>
- 7.20 Y.E. Bar-Sever: A new model for GPS yaw attitude, *J. Geod.* **70**(11), 714–723 (1996)
- 7.21 F. Dilssner: GPS IIF-1 satellite, antenna phase centre and attitude modelling, *Inside GNSS* **5**(6), 59–64 (2010)
- 7.22 J. Kouba: A simplified yaw-attitude model for eclipsing GPS satellites, *GPS Solutions* **13**(1), 1–12 (2009)
- 7.23 S.S. Russell, J.H. Schaibly: Control segment and user performance, *Navigation* **25**(2), 166–172 (1978)
- 7.24 J. Taylor: The GPS operational control system Kalman filter description and history, Proc. ION GNSS, Portland (ION, Virginia 2010) pp. 2329–2366
- 7.25 T. Creel, A.J. Dorsey, Ph.J. Mendicki, J. Little, R.G. Mach, B.A. Renfro: New, improved GPS – The legacy accuracy improvement initiative, *GPS World* **17**(3), 20–31 (2006)
- 7.26 W. Bertiger, Y. Bar-Sever, N. Harvey, K. Miller, L. Romans, J. Weiss, L. Doyle, T. Solorzano, J. Petzinger, A. Stell: Next generation GPS ground control segment (OCX) navigation design, Proc. ION GNSS, Portland (ION, Virginia 2010) pp. 964–977
- 7.27 W. Bertiger, Y. Bar-Sever, E. Bokor, M. Butala, A. Dorsey, J. Gross, N. Harvey, W. Lu, K. Miller, M. Miller, L. Romans, A. Sibthorpe, J. Weiss, M. Jones, J. Holden, A. Donigian, P. Saha: First orbit determination performance assessment for the OCX navigation software in an operational environment, Proc. ION GNSS, Nashville (ION, Virginia 2012)

- 7.28 P. Collins, R. Langley, J. LaMance: Limiting factors in tropospheric propagation delay error modelling for GPS airborne navigation, Proc. ION 52nd Annu. Meet., Cambridge (ION, Virginia 1996) pp. 519–528
- 7.29 D.D. McCarthy, G. Petit: *ERS Conventions (2003) IERS Technical Note No. 36* (des Bundesamts für Kartographie und Geodäsie, Frankfurt 2004)
- 7.30 B. Wiley, D. Craig, D. Manning, J. Novak, R. Taylor, L. Weingarth: NGA's role in GPS, Proc. ION GPS, Fort Worth (ION, Virginia 2006) pp. 2111–2119
- 7.31 C.H. Yinger, W.A. Feess, R. Di-Esposti, A. Chasko, B. Cosentino, B. Wilson, B. Wheaton: GPS satellite interfrequency biases, Proc. ION Annu. Meet., Cambridge (ION, Virginia 1999) pp. 347–354
- 7.32 Navstar GPS Control Segment to User Support Community Interfaces (Global Positioning Systems Directorate, California 2010) ICD-GPS-240A, 12 Jan. 2010
- 7.33 Navstar GPS Space Segment/Navigation User Segment Interfaces, Interface Specification (Global Positioning Systems Directorate, California 2013) IS-GPS-200H, 24 Sep. 2013
- 7.34 R. Gold: Optimal binary sequences for spread spectrum multiplexing, IEEE Trans. Inf. Theory **13**(4), 619–621 (1967)
- 7.35 B. Barker, J. Betz, J. Clark, J. Correia, J. Gillis, S. Lazar, K. Rehborn, J. Stratton: Overview of the GPS M code signal, Proc. ION NTM, Anaheim (ION, Virginia 2000) pp. 542–549
- 7.36 R.D. Fontana, W. Cheung, T. Stansell: The new L2 civil signal, GPS World **12**(9), 28–34 (2001)
- 7.37 A.J. Van-Dierendonck, C.J. Hegarty: The new L5 civil GPS signal, GPS World **11**(9), 64–72 (2000)
- 7.38 Navstar GPS Space Segment/User Segment L5 Interfaces, Interface Specification (Global Positioning Systems Directorate, California 2013) IS-GPS-705D, 24 Sep. 2013
- 7.39 J.J. Rushanan: The spreading and overlay codes for the L1C signal, Navigation **54**(1), 43–51 (2007)
- 7.40 F. Neuman, L. Hofman: New pulse sequences with desirable correlation properties, Proc. Natl. Telem. Conf. (1971)
- 7.41 J.W. Betz: Binary offset carrier modulations for radionavigation, Navigation **48**(4), 227–246 (2001)
- 7.42 H.L. van Trees: *Detection, Estimation, and Modulation Theory – Part 1* (John Wiley, New York 2001)
- 7.43 J.W. Betz, M.A. Blanco, C.R. Cahn, P.A. Dafesh, C.J. Hegarty, K.W. Hudnut, V. Kasemsri, R. Keegan, K. Kovach, L.S. Lenahan, H.H. Ma, J.J. Rushanan, D. Sklar, T.A. Stansell, C.C. Wang, S.K. Yi: Description of the L1C signal, Proc. ION GNSS (2006) pp. 2080–2209
- 7.44 Navstar GPS Space Segment/User Segment L1C Interfaces, Interface Specification (Global Positioning Systems Directorate, California 2013) IS-GPS-800D, 24 Sep. 2013
- 7.45 G.W. Hein, J.A. Avila-Rodriguez, S. Wallner, A.R. Pratt, J. Owen, J.L. Issler, J.W. Betz, C.J. Hegarty, S. Lt: Lenahan, J.J. Rushanan, A.L. Kraay, T.A. Stansell: MBOC: The new optimized spreading modulation recommended for Galileo L1 OS and GPS L1C, Inside GNSS **1**(4), 57–66 (2006)
- 7.46 T. Walter, J. Blanch: Characterization of GNSS clock and ephemeris errors to support ARAIM, Proc. ION PNT 2015, Honolulu (ION, Virginia 2015) pp. 920–931
- 7.47 S.T. Hutsell, G. Dieter, G. Hatten, T. Dass, J. Harvey: GPS clock/timescale management in the master control station, Proc. 35th Annu. PTI Meet., San Diego (2003)
- 7.48 S.T. Hutsell, B.K. Brottlund, C.A. Harris: How old is your GPS navigation message?, Proc. ION GPS, Salt Lake City (ION, Virginia 2000) pp. 2556–2561
- 7.49 J.A. Klobuchar: Ionospheric time–delay algorithm for single–frequency GPS users, IEEE Trans. Aerosp. Electron. Syst. **23**(3), 325–331 (1987)
- 7.50 R. Kenneth, Brown Jr.: The theory of the GPS composite clock, Proc. ION GPS, Albuquerque (ION, Virginia 1991) pp. 223–242
- 7.51 T.E. Parker, D. Matsakis: Time and frequency dissemination: Advances in GPS transfer techniques, GPS World **15**(11), 32–38 (2004), November
- 7.52 D.R. Hinson: *Letter to Dr. A. Kotaite* (Federal Aviation Administration, Washington 1994), Oct. 14
- 7.53 M.C. Blakey: *Letter to Dr. R. Kobeh* (Federal Aviation Administration, Washington 2007), Sep. 10
- 7.54 W.J. Clinton: *Statement by the President Regarding the United States Decision to Stop Degrading Global Positioning System Accuracy* (White House, Office of the Press Secretary, Washington D.C. 2000), May 1
- 7.55 D. Perino: *Statement by the Press Secretary* (White House, Office of the Press Secretary, Washington D.C. 2007), Sep. 18
- 7.56 K.T. Woo: Optimum semicodeless carrier–phase tracking of L2, Navigation **47**(2), 82–99 (2000)
- 7.57 US Department of Defense: *Preservation of Continuity for Semi–Codeless GPS Applications* (US Federal Register, Washington DC 2008), 23 September
- 7.58 2014 Federal Radionavigation Plan, (US Departments of Defense, Transportation, and Homeland Security, Washington D.C. 2015)

GLONASS

8. GLONASS

Sergey Revnivykh, Alexey Bolkunov, Alexander Serdyukov (deceased), Oliver Montenbruck

The Global'naya Navigatsionnaya Sputnikova Sistema (GLONASS) is a global navigation satellite system developed by the Russian Federation. Similar to its US counterpart, the NAVSTAR global positioning system (GPS), GLONASS provides dual-frequency L-band navigation signals for civil and military navigation. Initiated in the 1980s, the system first achieved its full operational capability in 1995. Following a temporary degradation, the nominal constellation of 24 satellites was ultimately reestablished in 2011 and the system has been in continued service since then. This chapter describes the architecture and operations of GLONASS and discusses its current performance. In addition, the planned evolution of the space and ground segment are outlined.

8.1 Overview	219
8.1.1 History and Evolution	219
8.1.2 Constellation	220
8.1.3 GLONASS Geodesy Reference PZ-90	221
8.1.4 GLONASS Time	223
8.2 Navigation Signals and Services	225
8.2.1 GLONASS Services	225
8.2.2 FDMA Signals	226
8.2.3 CDMA Signals	229
8.3 Satellites	232
8.3.1 GLONASS I/II	233
8.3.2 GLONASS-M	235
8.3.3 GLONASS-K	236
8.4 Launch Vehicles	237
8.5 Ground Segment	238
8.6 GLONASS Open Service Performance	241
References	243

8.1 Overview

Next to GPS, GLONASS is the second fully operational and global navigation satellite system. This section outlines the history of GLONASS and describes its basic characteristics.

8.1.1 History and Evolution

GLONASS is the second-generation dual-use governmental global satellite navigation system of the Russian Federation. The predecessor of GLONASS – the low-altitude satellite navigation/communication system Tsyklon/Tsikada – became operational in 1976 [8.1, 2]. It was available for civil users as well. The system provided a positioning accuracy of 80–100 m with a delay of 1.5–2 h. The idea of using the Doppler radio-frequency shift of signals transmitted from the satellites for navigation implemented in Tsyklon/Tsikada was introduced by Prof. Shebshayevich from the Mozhaysky Military Space Academy in 1957.

GLONASS research and development started in the beginning of the 1970s and was based on the method of instant position determination using measurements of time differences between a navigation receiver and a group of satellites emitting navigation signals with synchronized time tags. The first test satellite of the GLONASS system (named Uragan, or Hurricane, at that time) was successfully launched on 12 October 1982. On 24 September 1993, the GLONASS system with initial operational capability of 12 satellites was commissioned for military service. GLONASS with full operational capability (24 satellites) was deployed in 1995.

Civil use of GLONASS for air safety was first offered in 1988, when details of the GLONASS system and signal were released to the International Organization for Air Safety (ICAO) [8.3]. In parallel to these activities, various Western researchers had already made efforts to identify key properties of the

GLONASS signals through high-gain antenna measurements and a systematic code search [8.4–6]. These enabled early developments of standalone-GLONASS and GPS/GLONASS receivers. In 1995, through the decree of the President of the Russian Federation, GLONASS received the status of a dual-use system available for civil users worldwide. A comprehensive interface control document (ICD) describing the open service signals and navigation message was first made available in the English language in 1998 and has been continuously maintained since then [8.7].

Due to a limited operational lifetime of the early generation of GLONASS satellites and an insufficient replenishment, the number of active satellites gradually decreased down to a minimum of seven active satellites in 2001 (Fig. 8.1) [8.2, 8]. Since 2002, GLONASS sustainment and evolution have been conducted in the framework of the long-term GLONASS Federal Program with a secured budget enabling significant step-by-step performance improvements. Within the following decade, regular launches and the extended lifetime of the new GLONASS-M satellites helped to gradually increase the number of operational satellites. The 24-satellite constellation (Fig. 8.2) required for a fully global service was ultimately re-established in 2011 with the launch of GLONASS-M No. 44.

On 17 May 2007, the decree of the President of the Russian Federation declared the GLONASS open service available to all national and international users without any limitations [8.9]. At the same time, it was demanded that GLONASS-based navigation equipment shall be used by federal authorities for the sake of national security [8.10]. The GLONASS system is thus considered a core element in the implementation of the National Navigation Policy. Its maintenance and modernization in the 2012–2020 time frame are pursued

within the GLONASS Federal Program, which aims for a continued performance increase through improvements of both the ground and space system.

The value of GLONASS as a standalone or complementary system for geodesy and precise navigation was recognized early on by the scientific community [8.11] and promoted the buildup of global tracking networks with GLONASS-capable global navigation satellite system (GNSS) receivers. Within the International GLONASS Experiment (IGEX-98 [8.12]) precise orbit solutions were generated for the first time and transformation parameters between the various realizations of the terrestrial reference frame (PZ-90, WGS-84, etc.) could be established. This work was later continued in the International GLONASS Service (IGLOS) pilot project [8.13], which provided the starting point for high-precision GLONASS point positioning applications.

8.1.2 Constellation

The GLONASS space segment nominally consists of 24 operational satellites, evenly distributed over three orbital planes (Fig. 8.2). The nominal constellation parameters are summarized in Table 8.1.

For best coverage, GLONASS adopts a so-called Walker 24/3/1 constellation geometry, where the parameters $t/p/f$ specify the total number of satellites, the number of orbital planes and the phasing parameter [8.14]. The longitude of ascending node of each

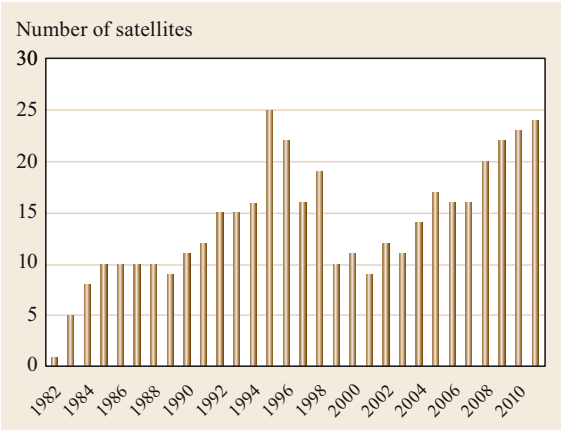


Fig. 8.1 GLONASS constellation development (after [8.8], courtesy of Springer)

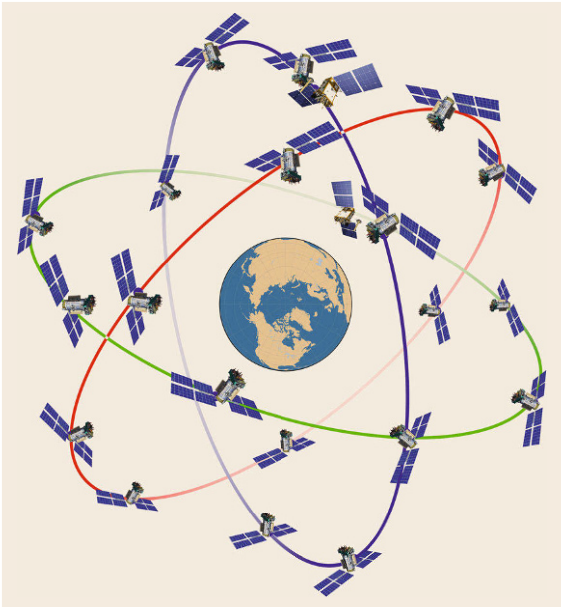


Fig. 8.2 GLONASS constellation in spring 2016 (courtesy of ISS Reshetnev)

Table 8.1 Nominal GLONASS constellation parameters

Parameter	Value
Number of operational satellites	$t = 24$
Number of orbital planes, p	$p = 3$
Number of satellites in a plane	$t/p = 8$
Phasing parameter	$f = 1$
Orbit type	Near circular
Eccentricity	$e < 0.01$
Inclination	$i = 64.8^\circ \pm 0.3^\circ$
Nominal altitude	$h = 19\,100\text{ km}$
Period of revolution	$T = 11\text{ h } 15\text{ min } 44\text{ s} \pm 5\text{ s}$
Longitude of ascending node between planes	$\Delta\Omega = 120^\circ$
Argument of latitude difference	$\Delta u = 45^\circ$
Latitude shift between planes	$\Delta u f / n = 15^\circ$
Ground track repeat cycle	17 orbits/8 d

plane differs by $\Delta\Omega = 360^\circ/p = 120^\circ$ from plane to plane. There are $t/p = 8$ satellites per plane, separated by $360^\circ p/t = 45^\circ$ in argument of latitude. The difference in the argument of latitude of satellites in equivalent slots in two different orbital planes is $\Delta u = 360^\circ f/t = 15^\circ$.

Each GLONASS satellite is identified by its *slot* number, which defines the orbital plane and its location within the plane (Fig. 8.3). Slot numbers 1–8 belong to orbital plane 1, while planes 2 and 3 comprise slot numbers 9–16 and 17–24 respectively.

The GLONASS satellites have no resonance with rotation of the Earth (based on gravitational field harmonics). The satellite's period is selected so that satellites make 17 full orbits for eight equinoctial days (approximately eight constituent days). Furthermore, the beginning of each orbit shifts with respect to the

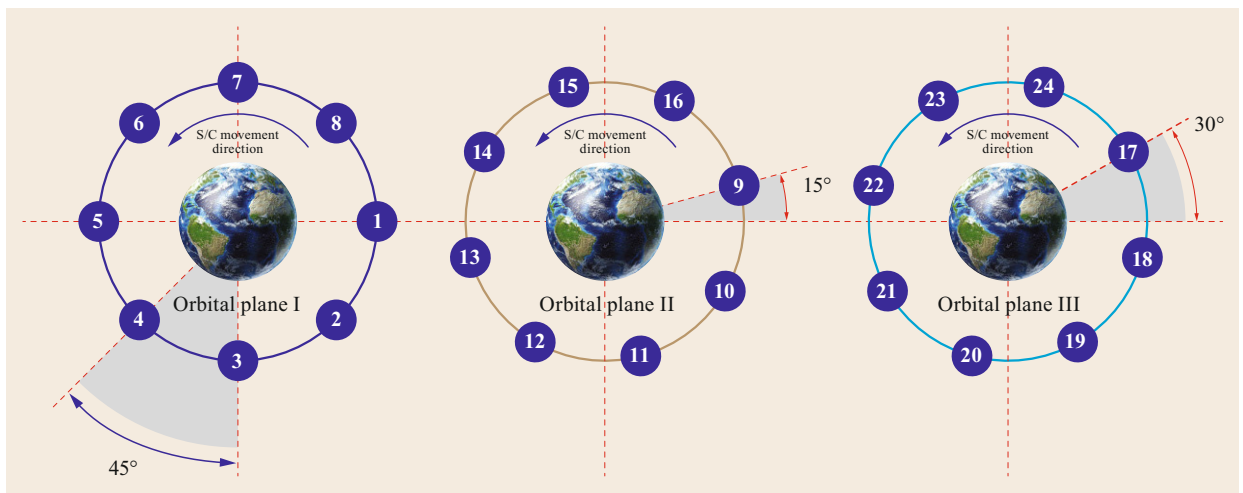
Earth's surface. Each eight days a satellite passes over the same point on the Earth's surface. Due to shifting in orbital planes, all the satellites are moving relative to Earth's surface practically along the same ground tracks (Fig. 8.4).

The orbital inclination of the GLONASS satellites ($\approx 65^\circ$) is roughly ten degrees higher than that of other medium altitude Earth orbit (MEO) navigation systems (GPS, BeiDou, Galileo), which provides improved visibility conditions over the area of the Russian Federation. Worldwide GLONASS users likewise benefit from a good sky coverage with a reduced visibility gap around the celestial pole (Fig. 3.8 of Chap. 3).

8.1.3 GLONASS Geodesy Reference PZ-90

The parameters and data of the Earth Model PZ-90 [8.15] are applied for GLONASS satellite orbit determination and ephemeris calculation. The PZ-90 system was established in 1990 and superseded the Soviet Geodetic System (SGS-85) that was used by GLONASS until 1993 [8.16].

The PZ-90 definition comprises fundamental geodetic constants, parameters of the Earth ellipsoid, and the Earth gravity field parameters (Table 8.2) as well as the geocentric reference system (GRS), which is defined in accord with common conventions of the International Earth Rotation and Reference Systems Service (IERS) and Bureau International De l'Heure (BIH). The PZ-90 system originates in the Earth's center of mass including the oceans and atmosphere. Its z -axis is directed to the conventional reference pole and its x -axis points to the intersection of the equatorial plane and the zero meridian as defined by the BIH [8.17, 18].

**Fig. 8.3** GLONASS satellite positions within orbital planes (courtesy of ISS Reshetnev)

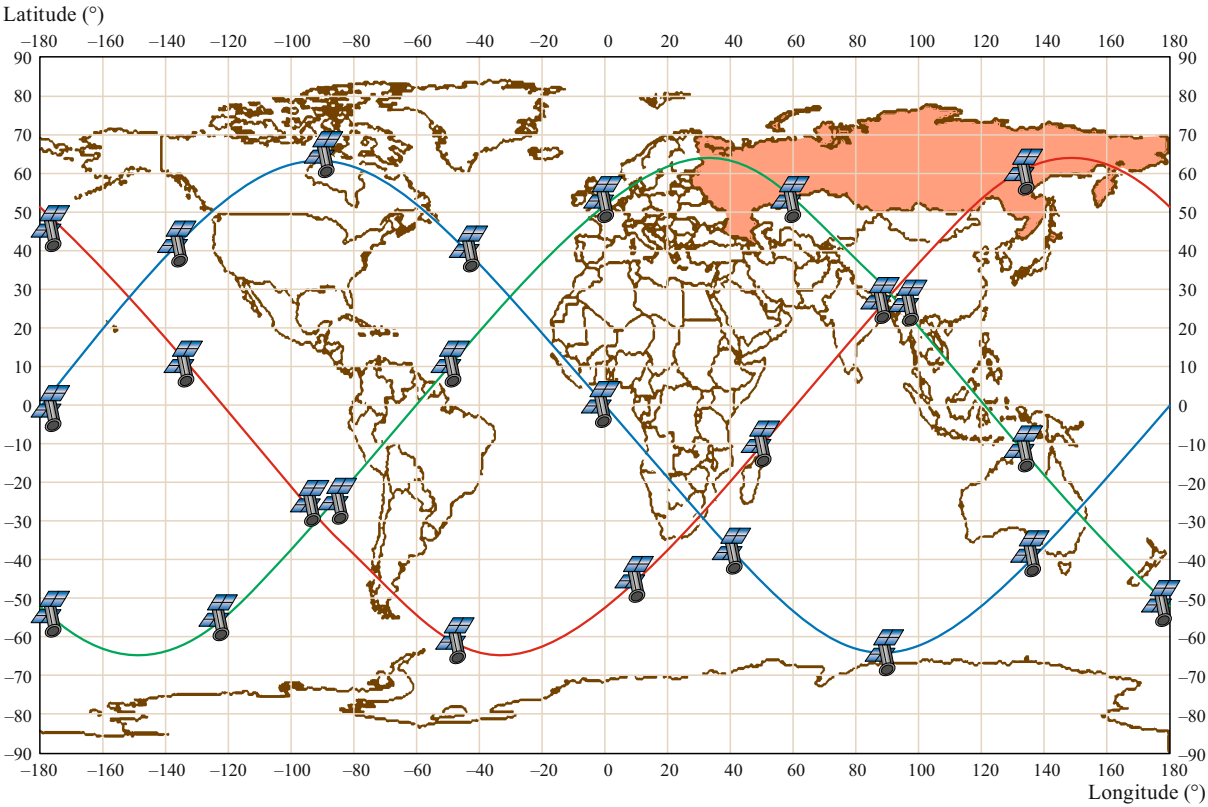


Fig. 8.4 GLONASS satellite ground tracks

Table 8.2 Fundamental parameters of the Earth model PZ-90 (after [8.15])

Parameter	Value
Speed of light in vacuum	$c = 299\,792\,458\text{ m/s}$
Gravitational constant	$G = 6.67259 \cdot 10^{-11}\text{ m}^3/(\text{kg s}^2)$
Geocentric gravitational coefficient (including atmosphere)	$GM_{\oplus} = 398\,600.4418 \cdot 10^9\text{ m}^3/\text{s}^2$
Angular velocity	$\omega_{\oplus} = 7.292115 \cdot 10^{-5}\text{ rad/s}$
Semimajor axis	$a = 6\,378\,136.0\text{ m}$
Flattening	$f = 1/298.257\,84$

The initial realization of the PZ-90 system had an accuracy of about 1–2 m. In the late 1990s, various efforts were made to establish the transformation between PZ-90 and the WGS-84 frame of GPS based on the joint processing of GPS/GLONASS observations in global networks and the comparison of postprocessed and broadcast GLONASS orbits [8.19–21].

The first major revision of the PZ-90 frame realization refers to the year 2002 and is known as PZ-90.02. It was introduced in GLONASS operations on 20 Septem-

ber 2007 [8.22] and notably improved the consistency of broadcast orbits with WGS-84 and the ITRF.

A further update, PZ-90.11 [8.15, 18] was implemented in the GLONASS operations on 31 December 2013 at 3:00 p.m. The PZ-90.11 GRS is a practical realization of the International Terrestrial Reference System (ITRS) at epoch 2010.0, which is based on the results of GPS/GLONASS data processing from space geodetic network (SGN) sites and a number of International GNSS Service (IGS) sites (Fig. 8.5). The accuracy (root mean square, RMS) of the PZ-90.11 GRS with respect to the Earth center is 0.05 m with relative accuracy of the reference point position on the level of 0.001–0.005 m.

The transition between different frames is commonly described by a seven-parameter similarity transformation (*Helmert transformation*)

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}_B = \begin{pmatrix} 1+m & +\omega_z & -\omega_y \\ -\omega_z & 1+m & +\omega_x \\ +\omega_y & -\omega_x & 1+m \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}_A + \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} \tag{8.1}$$

for the coordinates $\mathbf{r}_A = (x, y, z)_A^\top$ and $\mathbf{r}_B = (x, y, z)_B^\top$ in the original (A) and transformed (B) system, where



Fig. 8.5 PZ-90.11 reference points on the Russian territory (as of 2011)

Table 8.3 PZ-90 transformation parameters

From	To	ΔX (m)	ΔY (m)	ΔZ (m)	ω_x ($10^{-3}''$)	ω_y ($10^{-3}''$)	ω_z ($10^{-3}''$)	m (10^{-6})	Epoch	Reference
PZ-90	WGS-84	-1.10	-0.30	-0.90	0	0	-200	-0.12	1990.0	[8.17, 23]
PZ-90	ITRF-97	+0.07	+0.00	-0.77	-19	-4	+353	-0.003		[8.21]
PZ-90	PZ-90.02	-1.07	-0.03	+0.02	0	0	-130	-0.22	2002.0	[8.17, 23]
PZ-90.02	WGS-84(1150)/ ITRF-2000	-0.36	+0.08	+0.18	0	0	0	0	2002.0	[8.17, 23]
PZ-90.11	ITRF-2008	-0.003	-0.001	+0.000	+0.019	-0.042	+0.002	-0.000	2010.0	[8.18]

$\Delta \mathbf{r}_B = (\Delta x, \Delta y, \Delta z)^\top$ is the translational offset, $\boldsymbol{\omega} = (\omega_x, \omega_y, \omega_z)^\top$ describes the rotational transformation and m denotes the scale difference. Transformation parameters for past and current realizations of the PZ-90, WGS84, and the International Terrestrial Reference Frame (ITRFs) are summarized in Table 8.3.

8.1.4 GLONASS Time

GLONASS System Time (GLST) provides the common reference to which all GLONASS satellite clocks are synchronized. It is based on observations of an ensemble of continuously operating hydrogen masers in the GLONASS ground segment and synchronized to Universal Time Coordinated of Russia, UTC(SU) as a reference timescale. UTC(SU) is itself maintained by the National Metrology Institute of the Russian Federation (VNIIFTRI) [8.24] in Mendeleevo near Moscow as part of the State Time and Frequency Service (STFS). It is realized through an ensemble of hydrogen masers and

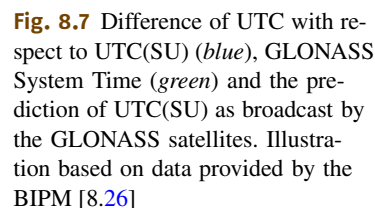
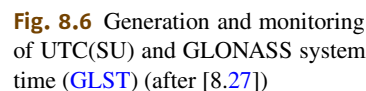
continuously steered to Coordinated Universal Time (UTC) through satellite time and frequency transfer techniques [8.25]. Differences of UTC and UTC(SU) are routinely monitored by the Bureau International des Poids et Mesures (BIPM) and have decreased from a few tens of ns in 2011 [8.26] to less than 2 ns in 2016.

The comparison of the national timescale, GLST, and UTC is accomplished through common-view GNSS time transfer (using either GPS or GLONASS satellites; Chap. 41) and two-way time and frequency transfer via geostationary satellites (Fig. 8.6).

Unlike GPS time, the GLONASS System Time exhibits no integer offset from UTC, but is offset by 3 h to match the local time zone of Moscow as adopted by the GLONASS ground control segment

$$\text{GLST} = \text{UTC}(\text{SU}) + 3 \text{ h} - C. \quad (8.2)$$

The fractional difference C is controlled to be less than $1 \mu\text{s}$ [8.7]. A predicted value thereof, the GLONASS



board timescales and the system timescale. The frequency and time corrections are calculated for each orbit and uploaded on board the satellites for transmission to the users. The frequency and time corrections are two parameters of the linear approximation of the onboard timescale offset relative to the system timescale.

Within the frequency and timekeeping facility of the GLONASS *central synchronizer* (CS), the frequencies and phases of four hydrogen frequency standards are continuously compared and the best clock is used as primary standard. The resulting 5 MHz signal exhibits a frequency error of less than $3 \cdot 10^{-14}$ and a stability of better than $2 \cdot 10^{-15}$ over one day. For redundancy purposes, an independent master and backup CS system are operated at Schelkovo (near Moscow) and Komsomolsk, respectively. Monitoring of the CS time with respect to UTC(SU) is performed through common-view time transfer between the CS and STFS. De-

The GLONASS time synchronization system of the ground control segment generates the system timescale, calculates frequency and time corrections, determines the difference between the system timescale and UTC(SU), and calculates corrections between on-

pending on the employed equipment and signals (GPS or GLONASS), a monitoring precision of 3–13 ns is achieved.

The fractional differences between UTC and GLST as well as that of UTC and the prediction $UTC(SU)_{GLO}$ broadcast by the GLONASS satellites are routinely monitored by the BIPM and published as part of the monthly *Circular T*. The values presently demonstrate a stability of better than 10 ns, but are potentially af-

fected by systematic offsets at the level of few hundreds of ns due to overall calibration uncertainties (Fig. 8.7). To improve the alignment of GLONASS System Time and the predicted UTC(SU), various adjustments have been performed starting on 18 August 2014. These included corrected offsets for the broadcast time parameters and a gradual tuning of the ground clocks [8.28]. As a result of the alignment, GLST and $UTC(SU)_{GLO}$ have exhibited a consistency of about 10 ns since early 2015.

8.2 Navigation Signals and Services

8.2.1 GLONASS Services

In accord with its dual-use status, GLONASS provides two types of services:

- An open service with unencrypted signals in up to three frequency bands (L1, L2, and recently L3) that is globally available for all users without any limitations
- A service for authorized users, using encrypted signals in presently two frequency bands (L1, L2).

The terminology of these services is not well defined, however, and alternative expressions such as *standard positioning service* and *high-precision service* are commonly used in the open literature.

Performance specifications for the two service types have not been released so far, but a GLONASS Open Service Performance Parameters Standard is under discussion within the International Committee on GNSS (ICG). Initial drafts suggest a performance specification of about 5 and 10 m, respectively for the 95% (2σ) global average position error in horizontal and vertical directions [8.29]. The option of an intentional accuracy degradation of the open service (similar to the *Selective Availability* employed by GPS up to the year 2000) has never been considered in the GLONASS system design.

The authorized service is primarily intended for military users [8.7]:

The PP [pinpoint accuracy] signal is modulated by a special code and intended for usage in interests of the Ministry of Defense. Usage of a PP signal should be agreed with the Russian Federation Defense Ministry.

Unlike GPS, the signals of the authorized service are presently not encrypted. Even though their structure and data contents have not been publicly released by the system providers, the employed ranging code has, nevertheless, been revealed already in the early days of GLONASS through a systematic code search [8.5].

This has enabled the design of geodetic dual-frequency GLONASS receivers and allowed for an early use of GLONASS in precise point positioning applications. In accord with the above disclaimer, access to these signals may, however, be inhibited and its *unofficial* use should be considered with due care.

Each GLONASS satellite transmits signals for both the open and the authorized service. The L1/L2 signals for these two services are designated as standard-accuracy (ST) and high-precision (or pinpoint, PP) signals in the GLONASS ICD [8.7], to reflect the different performance of the employed ranging codes.

Traditionally, GLONASS makes use of frequency division multiply access (FDMA) modulation. Here, signals transmitted by individual satellites employ the same ranging code, but slightly different frequencies to allow concurrent processing in the receiver. With the first *GLONASS-K1* satellite launched in 2011, GLONASS started to transmit additional code division multiple access modulation (CDMA) signals on the new L3 signal. As part of the ongoing GLONASS modernization, CDMA signals will also be transmitted in the L1 and L2 bands to improve interoperability with other GNSSs, specifically GPS. An overview of current and planned signals is given in Table 8.4.

Table 8.4 GLONASS signals overview. Signals are identified by the frequency band (first two characters), the service type (O: open, S: authorized special), and the modulation type (F: FDMA, C: CDMA)

Satellites	FDMA		CDMA		
	L1	L2	L1	L2	L3
GLONASS	L1OF L1SF	L2SF			
GLONASS-M	L1OF L1SF	L2OF L2SF			(L3OC)
GLONASS-K1	L1OF L1SF	L2OF L2SF			L3OC
GLONASS-K2	L1OF L1SF	L2OF L2SF	L1OC L1SC	L2OC L2SC	L3OC

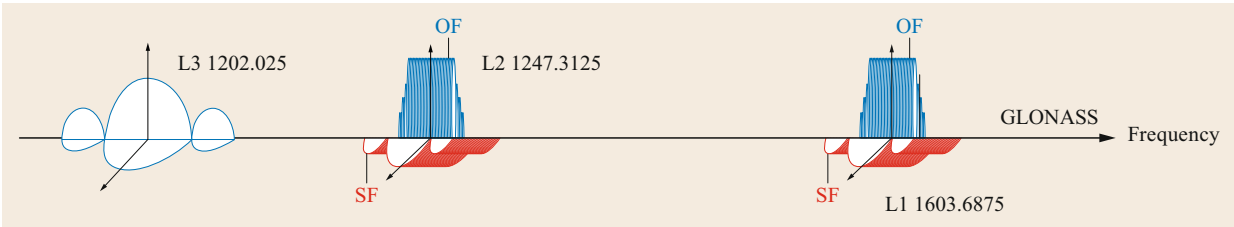


Fig. 8.8 GLONASS frequency division multiple access (FDMA) and code division multiple access (CDMA) signals in the L1, L2 and L3 band (status 2015)

The spectral distribution of currently transmitted signals in the L1, L2, and L3 bands is illustrated in Fig. 8.8. The L1 and L2 FDMA signals employ slightly higher (≈ 20 MHz) frequencies than the corresponding GPS signals. The GLONASS L3 signal, in contrast, is transmitted at a widely different frequency ($f_{L3} = 1202.025$ MHz) than the L3 frequency allocated to the GPS nuclear detection (NUDET) payload ($f_{L3} = 1381.05$ MHz). In fact, the GLONASS L3 frequency closely matches the Galileo E5b frequency with only a small negative offset.

For the provision of positioning, navigation and timing services, the GLONASS signals include navigation messages with ephemeris data in the PZ90 reference system and timing parameters related to GLONASS System Time.

The open service signals as well as the contents of the navigation message are detailed in the public signal ICD [8.7], which presently covers the L1/L2 FDMA signals and will be updated with L3 and CDMA signals as soon as these become fully operational.

8.2.2 FDMA Signals

GLONASS has been transmitting FDMA signals since the first satellite was launched. Despite the ongoing system evolution and the introduction of new CDMA

signals, GLONASS will continue to provide the legacy FDMA signals in the future to provide backward compatibility with the user equipment already in use.

FDMA signals are transmitted in the L1 and L2 bands and comprise the open service (standard-accuracy) and authorized service (high-precision) ranging codes (Table 8.5). In analogy with GPS, the two codes are commonly described as GLONASS C/A-code (coarse and acquisition code) and P-code (precise code) respectively, even though these terms are not mentioned in the ICD and do not represent an official designation.

The open service signal was only transmitted on the L1 frequency in the first generation of GLONASS satellites, but has been made available on both frequencies starting with the GLONASS-M series of spacecraft. Also, the L2 signal power, which was initially about 6 dB lower, has been adjusted to the same level as L1.

Signal Frequencies

GLONASS FDMA signals use a distinct set of channels for different signals. Each channel is identified by its channel number k , which uniquely defines the corresponding signal frequency

$$\begin{aligned} f_{L1}(k) &= (1602.0 + k \cdot 0.5625) \text{ MHz}, \\ f_{L2}(k) &= (1246.0 + k \cdot 0.4375) \text{ MHz}. \end{aligned} \quad (8.3)$$

Table 8.5 Main characteristics of the legacy GLONASS FDMA signals. Parameters of the authorized service signals (L1SF, L2SF) that have not been publicly released are marked as N/A (not available)

Signal	Received power (dBW)	Center frequency (MHz)	Code and Data	Modulation	Bandwidth (MHz)	Data rate (bps)
L1OF	-161	1598.0625 ... 1605.375	C/A-code (511 chips, 1 ms) Open service navigation message	BPSK(0.5)	$\approx \pm 0.5$	50
L1SF	-161	same	P-code (5.11 MHz) Authorized service navigation message	N/A	$\approx \pm 5$	N/A
L2OF	-161	1242.9375 ... 1248.625	C/A-code (511 chips, 1 ms) Open service navigation message	BPSK(0.5)	$\approx \pm 0.5$	50
L2SF	-161	Same	P-code (5.11 MHz) Authorized service navigation message	N/A	$\approx \pm 5$	N/A

Neighboring channels are separated by $\Delta f \approx 0.5$ MHz, which roughly corresponds to the half-width of the open service signal spectral and is sufficient to distinguish transmissions from different satellites in the receiver. For a given channel number k , the ratio of the L1 and L2 frequencies attains a fixed value of

$$\frac{f_{L1}(k)}{f_{L2}(k)} = \frac{9}{7}. \quad (8.4)$$

In the original GLONASS design, use of a unique frequency channel number in the range of $k = 0, \dots, 24$ (corresponding to an L1 frequency in the range of 1602.0–1615.5 MHz) was considered for each individual spacecraft in the 24-satellite constellation along with a spare channel for testing purposes [8.30].

The initial frequency allocation was later modified in compliance with recommendation RA 769 of the International Telecommunication Union (ITU) on protection criteria used for radio astronomical observations [8.31], when it became obvious that GLONASS L1 transmissions interfered with observations of the hydroxyl radical (OH) near 1612 MHz [8.32]. Starting in 1998, the frequency indices were restricted to $k = 0, \dots, 12$ (yielding a maximum center frequency of 1608.75 MHz). In a second update, conducted in 2005, negative channel numbers were introduced and the covered range was changed to $k = -7, \dots, +6$ [8.7] (including two channels commonly used for testing new satellites). In addition, the GLONASS satellites were equipped with bandpass filters to reduce transmissions in the frequency band relevant for radio astronomy. This can readily be seen from the L1 signal spectrum of the GLONASS satellites, which exhibits a sharp gap around 1612 MHz (Fig. 8.9).

The limitation of only 12–14 channels for 24 satellites is handled by assigning identical frequency channel numbers to antipodal satellites, that is satellites in

opposite slots of the same orbital plane. Such satellites should never be jointly visible for a terrestrial observer and can therefore make use of the same signal frequency.

The use of different signal frequencies in the GLONASS FDMA signals enables use of a common ranging code for all satellites and offers protection against narrow-band interference compared to CDMA signals, since such interferences would only affect one or a few satellites at a time [8.33]. However, it also results in an increased complexity of the front-end design and is commonly a source of undesired group and phase delay variations in GLONASS receivers (Chap. 13). Nevertheless, the existing data processing methodology allows the detection of these delays at the user level providing precise point positioning computations.

Open Service Signal

The FDMA open service signal transmitted on the L1 and L2 frequencies employs a binary phase-shift keying (BPSK) modulation. The carrier is modulated with a binary sequence, which results from the modulo-2 addition (i.e., exclusive or combination) of three individual components:

- The pseudorandom noise (PRN) ranging code
- The navigation data
- An auxiliary meander sequence.

The ranging code has a length of 511 chips and is clocked at 511 MHz, thus yielding a total duration of 1 ms. Data bits are transmitted at a 50 Hz rate and have a length of 20 ms per data bit. The third component is a 100 Hz sequence of alternating 1 s and 0 s with a duration of 10 ms per symbol. It is termed a *meander* sequence or *Manchester code* and ensures that there is at least one transition in the modulo-2 sum of the meander and data sequence within each data bit interval. The

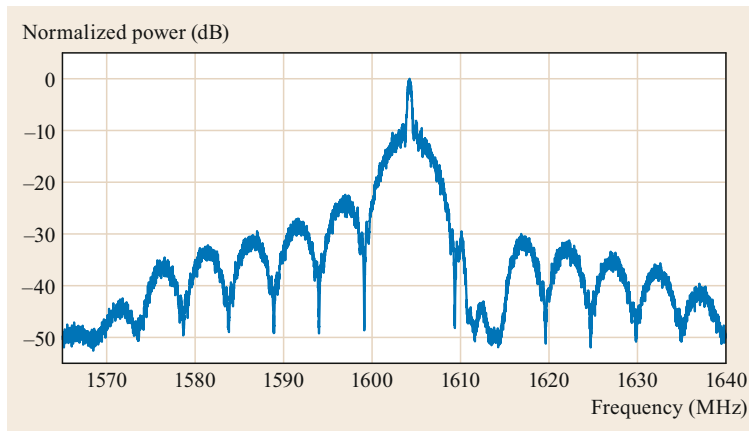


Fig. 8.9 Example of GLONASS L1 signal spectrum. The center frequency of 1604.25 MHz corresponds to a frequency channel number $k = +4$. The central peak results from the 511 kHz open service coarse and acquisition (C/A) code, while the wider lobes reflect the 5.11 MHz P-code modulation. Transmissions near 1612 MHz are masked by a bandpass filter to protect radio astronomical observations of hydroxyl (OH) emissions (courtesy of S. Thörlert, Deutsches Zentrum für Luft- und Raumfahrt (DLR))

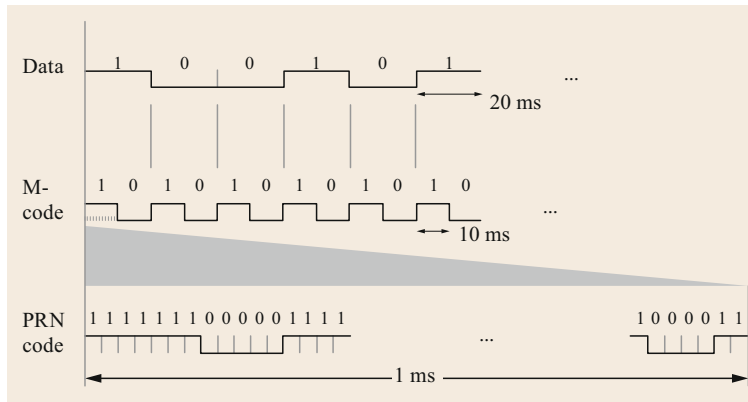


Fig. 8.10 Structure of the GLONASS open service signal

three signal components are synchronized to each other, that is there are 10 full ranging codes within a Manchester code symbol and 20 codes within a data bit. The overall signal structure is illustrated in Fig. 8.10.

The FDMA concept allows for use of a common PRN code by all GLONASS satellites and the same code is in fact also used on the two frequencies. Unlike the Gold codes of GPS, the GLONASS ranging code is obtained from just a single maximum-length linear feedback shift register (LFSR). Gold codes, in contrast, require a combination of two shift registers, from which a large family of PRN sequences with good cross-correlation properties can be constructed. This is important for use of CDMA signals in a GNSS constellation such as GPS, since distinct, high-quality PRN sequences are needed for the various spacecraft [8.34]. GLONASS, in contrast requires only a single ranging code and achieves a low cross-correlation between different satellites through the frequency separation of different transmission channels [8.33].

A more simple PRN code generator based on a 9 bit register can therefore be employed, which provides a maximum length pseudorandom sequence of 511 ($= 2^9 - 1$) chips (Fig. 8.11). The register is initialized with all 1s and the PRN ranging code is extracted at the output of the seventh stage of the shift register (Fig. 8.11).

The chipping rate and duration of the GLONASS ranging code are directly reflected in the corresponding

open service signal spectrum [8.4]. It exhibits an overall bandwidth of about 1 MHz with spectral nulls at multiples of 511 kHz relative to the center frequency and individual lines spaced at 1 kHz (i. e., the inverse of the code length). It may be noted that the bandwidth is larger than the spacing of individual frequency channels on both the L1 and L2 frequencies. This does not, however, inhibit a safe acquisition and tracking of the GLONASS signals, since the correlation bandwidth is substantially smaller (about 1 kHz for a 1 ms integration interval).

Authorized Service Signal

Both the L1 and L2 frequencies carry an additional authorized service signal, which is transmitted in phase quadrature to the open service signal. The authorized signal is intended for military use only, and its signal structure has never been publicly disclosed by official sources.

Nevertheless, it has been possible to reveal important properties already in the early days of GLONASS from inspection with a high-gain antenna and a systematic testing of different hypotheses on the code design. This work has mainly been conducted by the University of Leeds [8.4, 5, 35, 36] and forms the basis of today's understanding of the military signal and its implementation in geodetic receivers. Knowledge of the GLONASS P-code was of particular interest prior to the GLONASS-M satellite generation, since it provided access to the signal second frequency and thus enabled ionospheric correction of GLONASS observations in precise positioning applications. Also, the higher chipping rate offered a somewhat better noise and multipath performance [8.36] than the C/A-code. It must be kept in mind, though, that the P-code signal is not intended for public use and may change without prior notice.

Navigation Message Structure

GLONASS navigation messages for the FDMA open service use a fixed structure made up of individual

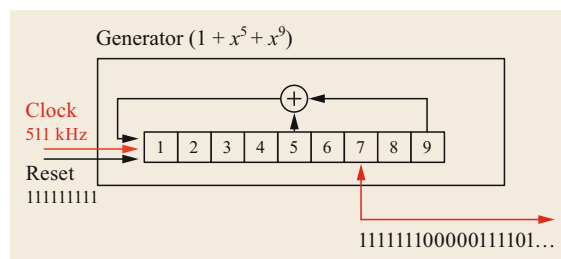


Fig. 8.11 GLONASS C/A-code generator

frames and *strings*. The entire set of frames is designated a *superframe* and repeated at regular intervals (Table 8.6).

The structure of the FDMA open service navigation message is illustrated in Fig. 8.12 based on information in [8.7]. Each string is made up of 85 bit and a time mark. This time mark consists of 30 symbols (corresponding to a hexadecimal value of 0x3E375096 [8.36]) and serves as a postamble. At a bit length of 20 ms and a symbol length of 10 ms, the entire string is transmitted in $1.7 \text{ s} + 0.3 \text{ s} = 2 \text{ s}$. The 85 bit comprise of a zero bit, the 4 bit string number, 72 bit of navigation data and 8 Hamming code parity bits offering single-error correction [8.36].

Strings 1–4 of each frame provide *immediate* (ephemeris) data of the transmitting satellite, which are required for the positioning and repeated once every 30 s. The remaining strings 4–15 contain sub-commutated *nonimmediate* (almanac) data for up to five satellites. The remaining bytes in the fifth frame hold additional parameters for conversion from GLONASS System Time to UT1 as well as the announcement of leap-second adjustments.

The entire superframe is transmitted in 2.5 min and continuously repeated between ephemeris updates (nominally once every 30 min). The overall message structure is adapted to support the current 24-satellite constellations but leaves some spare bytes for adding additional information.

Unlike GPS, the GLONASS system does not make use of an orbital elements representation of the ephemeris data, but provides the state vector (position and velocity) along with corrections in the Earth-fixed PZ90 coordinate system at the given reference epoch. Based on this information, the trajectory in the vicinity of this epoch can be obtained by numerical integration of the equation of motion (Sect. 3.3.3). To simplify the complexity of the orbit model, only the dominating Earth oblateness term is explicitly considered. However, additional acceleration corrections accounting for the effect of lunisolar perturbations are provided as part of the navigation message for improved accuracy. The

ephemeris data are typically updated at half-hourly intervals (i. e., at full and half hours) and are valid for $\pm 15 \text{ min}$ around the reference epoch in the center of the interval. Satellite clock offsets of the satellite with respect to GLST are described through a linear clock polynomial, which directly yields the apparent clock with no need for application of a relativistic correction. Furthermore, the immediate data of the GLONASS navigation message comprise health information, GLST to UTC time conversion data and a timing group delay parameter for consideration of differential code biases in single-frequency positioning.

The *nonimmediate* data of the FDMA open service navigation message provide timescale information (for GLST to UTC(SU) and GLST to GPS time conversion), which is repeated in string 5 of each frame as well as two strings of almanac data per satellite. Aside from orbital elements and a coarse clock offset value, the almanac includes health information as well as the slot and frequency channel number of the respective satellite.

Descriptions of the FDMA authorized navigation message as inferred from the analysis of transmitted P-code data in the late 1990s are provided in [8.35, 36].

It may be noted that the GLONASS FDMA navigation messages provide ephemeris, almanac and time system information but no ionospheric correction data for single-frequency users. For best positioning accuracy, an ionospheric compensation using dual-frequency L1/L2 observations is therefore recommended.

8.2.3 CDMA Signals

As part of the GLONASS signal evolution plan, new CDMA signals are made available to the users as a complement to the legacy FDMA signals. Key reasons for introducing CDMA signals include an improved navigation accuracy, an improved resistance to interference, and the improved separation of open and authorized services. The GLONASS signal evolution plan is based on a phased approach. A first CDMA signal in the L3 band has been made available since 2011 and further signals will be added with each new generation of GLONASS satellites. An initial assessment of GLONASS CDMA L3 ambiguity resolution and positioning performance is provided in [8.37].

Similar to FDMA signals, there are two types of GLONASS CDMA signals: open and encrypted, providing public and authorized services. The frequency allocations for L1 and L2 CDMA signals are defined within the original GLONASS frequency bands (Fig. 8.13), while L3 is a newly allocated frequency next to the Galileo E5b and BeiDou B2 band. In addition, the provision of modernized civil navigation signals (L1OCM, L5OCM) on the L1/E1 and L5/E5a

Table 8.6 Navigation data structure of open service GLONASS FDMA signals. HC and TM denote the 8 bytes of the Hamming error correction code and the 30 symbols of the time mark

Structure	Duration	Elements
Superframe	2.5 min	5 frames
Frame	30 s	15 strings
String	2 s	85 bit + timemark
Bit	0.02 s	–
Timemark	0.3 s	30 symbols
Symbol	0.01 s	–

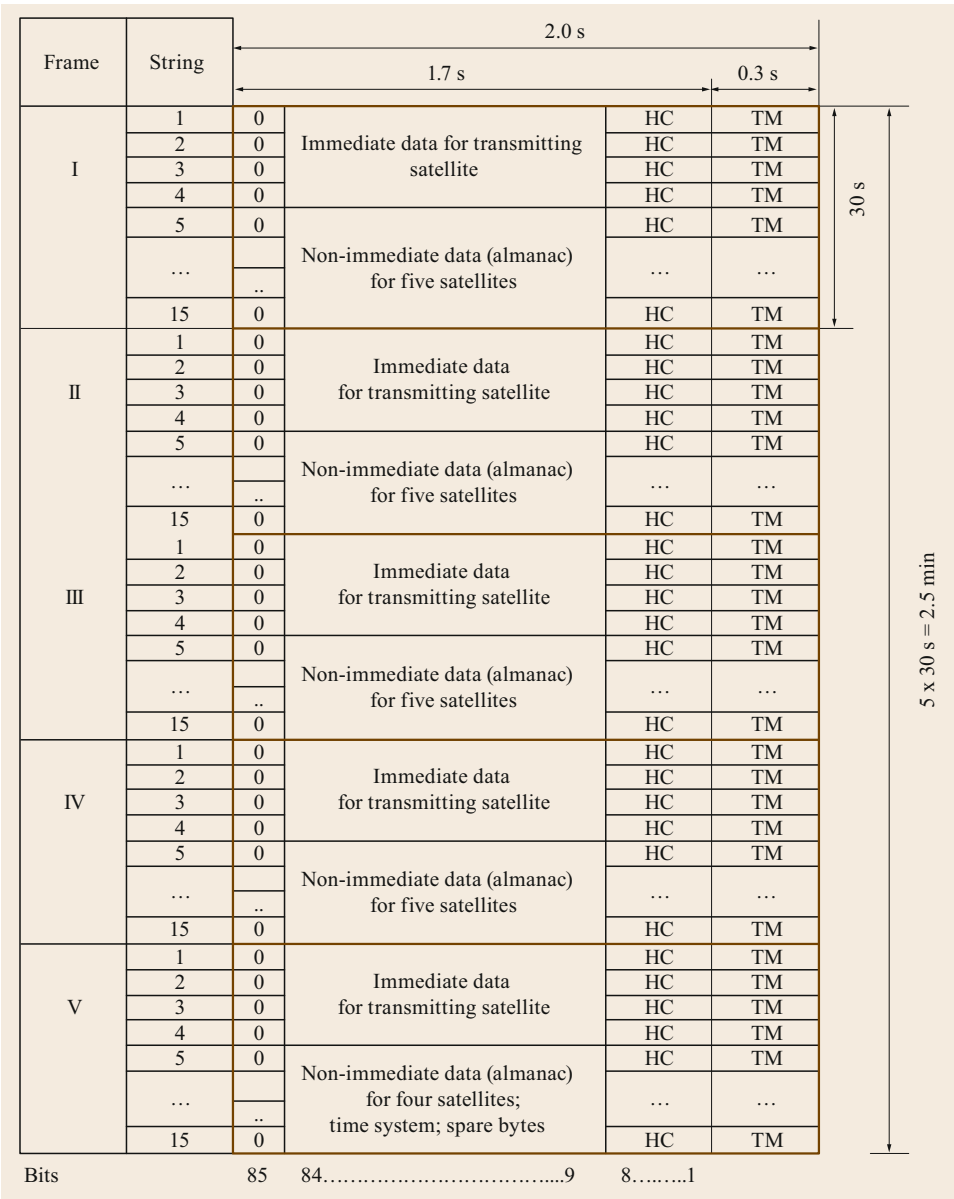


Fig. 8.12 Navigation message superframe structure for FDMA open service signals

frequencies of the GPS and Galileo system are under study to achieve a maximum compatibility with these other constellations.

The main parameters of current and proposed GLONASS CDMA signals are presented in Table 8.7 [8.8, 38, 39]. The open service L1OC and L2OC signals are planned to provide time-multiplexed data and pilot components using BPSK(1) and BOC(1,1) modulations respectively [8.8]. The L1SC and L2SC authorized service signals, in contrast, will use BOC(5,2.5) modulation for both the pilot and data components and are transmitted in phase quadrature

relative to the open signals. The BOC(5,2.5) modulation offers a good spectral separation of open and authorized signals, while simultaneously suppressing emissions in the radio astronomical frequency band around 1612 MHz [8.38].

A first CDMA signal, namely the L3OC open service signal, was introduced by the GLONASS-K1 satellite launched in 2011 and is also made available by the latest version of GLONASS-M satellites launched since 2014. While a formal ICD is pending, basic information on the signal structure has been publicly released in [8.38]. The signal comprises a data and

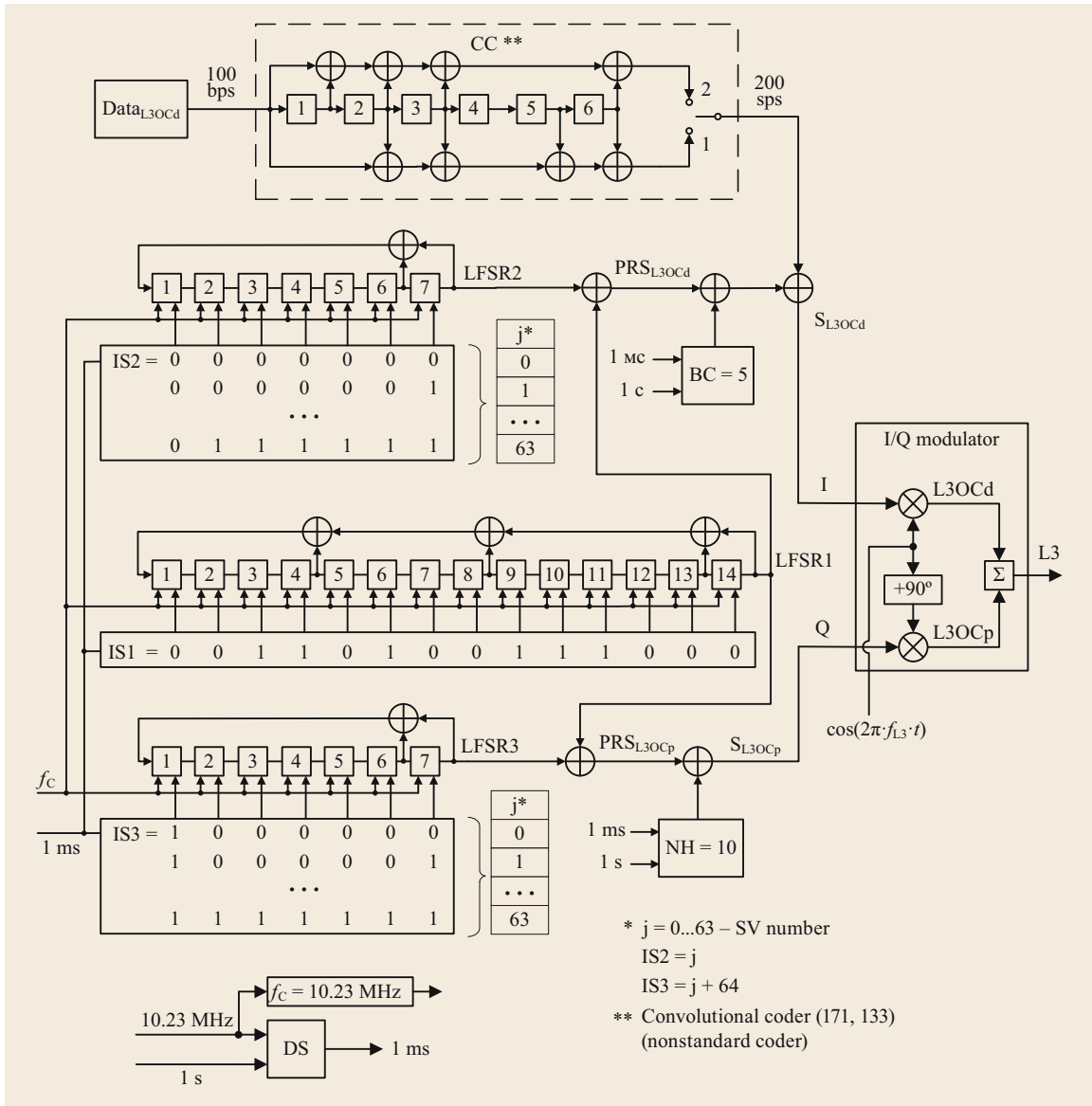


Fig. 8.14 GLONASS L3 open service CDMA signal generation (courtesy of Russian Space Systems Corporation)

a pilot component in phase quadrature using BPSK(10) modulation.

As illustrated in Fig. 8.14, the pseudorandom code for both components is generated from the modulo-2 addition of the outputs of a 14 bit shift register (IS2; with feedback taps 4, 8, 13, and 14) and a seven-stage shift register (IS1/IS3 for the pilot/data code; with feedback taps 6 and 7). While the initial state of the IS2 register is common to all spacecraft, satellite-dependent initial values are used for the IS1 and IS3 registers. The resulting PRN sequences are known as *Kasami* sequences [8.40] and exhibit a very low cross correla-

tion for an entire family of individual codes [8.41]. The L3OC codes have a native length of $2^{14} - 1 = 16\,383$ bit, but are truncated at 10 230 chips and achieve a cross correlation of -40 dB. At the employed chipping rate of 10.23 MHz, the primary code duration amounts to 1 ms.

Navigation data are modulated at a rate of 100 bps using a 1/2 convolutional encoding for error correction (yielding a symbol rate of 200 sps). In addition to the ranging code and the encoded navigation data, the data channel is modulated with a 5-bit Barker code at a rate of 1 kHz. The Barker code (BC) is synchronized with

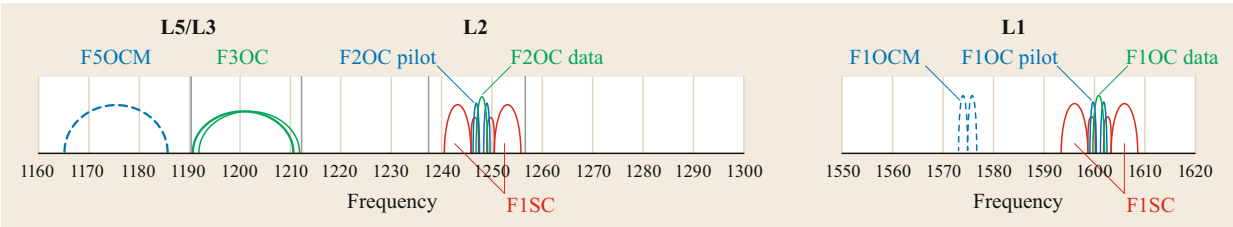


Fig. 8.13 GLONASS CDMA signal frequency allocations

Table 8.7 GLONASS CDMA signal parameters. Parameters of the authorized service signals that are not publicly disclosed are marked as N/A (not available). Parameters of signals currently under study are marked as TBD (to be defined)

Band Signal	L5/L3		L2		L1		
	L5OCM	L3OC	L2SC	L2OC	L1OCM	L1SC	L1OC
Access	Open	Open	Authorized	Open	Open	Authorized	Open
Carrier frequency (MHz)	1176.45	1202.025	1248.06		1575.42	1600.995	
Data signal modulation	BPSK(10)	BPSK(10)	BOC(5,2.5)	BPSK(1)	TBD	BOC(5,2.5)	BPSK(1)
Pilot signal modulation	BPSK(10)	BPSK(10)	BOC(5,2.5)	BOC(1,1)	TBD	BOC(5,2.5)	BOC(1,1)
Data rate (bps)	TBD	100	N/A	250	TBD	N/A	125
Navigation data (ms)	TBD	10	N/A	4	TBD	N/A	8
Chip rate rate (MHz)	10.23	10.23	N/A	0.5115	TBD	N/A	0.5115
Status	study	implementation			study	implementation	

the ranging sequence and covers the duration of a single navigation data symbol (i. e., 5 ms). For the pilot channel, a secondary (or Neuman–Hofman, [NH](#)) code with a length of 10 bit is employed. This yields an effective code length of 10 ms and offers increased robustness and weak-signal tracking capabilities. A detailed characterization of the L3OC signal as transmitted by the GLONASS-K1-1 satellite is provided in [8.42] based on analyses with a high-gain antenna and a software receiver.

Along with the advanced modulation scheme, the L3 CDMA signal also introduces a new navigation message concept. Unlike the fixed superframe structure of the FDMA navigation message, the L3OC signal uses a flexible message system [8.38, 39]. Similar to the [CNAV](#) (civil navigation) message of the GPS L2C and L5 signals, a set of distinct messages is defined, each of which provides a specific subset of

navigation data. Each message is uniquely identified by its message number and carries a cyclic redundancy check ([CRC](#)) field for error protection. Information from different messages is combined at user level to obtain the full set of navigation data. Since unknown messages types are specified to be ignored by a receiver, the new scheme greatly facilitates the addition of new messages along with signal upgrades and service improvements. Among others, the new navigation message is not limited to a predetermined size of the constellation but can accommodate a variable number of satellites.

By way of example, the layout of the new L3OC almanac message is described in [8.43]. Each message has a length of 300 bit including a 20 bit time mark and a 24 bit CRC. A full description of all L3OC navigation messages will be provided as part of the GLONASS open service signal ICD.

8.3 Satellites

The constellation of GLONASS satellites is a key element of the entire GLONASS system. Throughout the more than 30 years of its history, three generations of GLONASS satellites have been built and operated:

- The initial generation of *GLONASS* satellites first launched in 1982
- The subsequent *GLONASS-M* satellites launched since 2003

- The *GLONASS-K* series introduced in 2011.
- Each new generation of GLONASS satellites extended the satellite capabilities and improved the overall system performance. In parallel to technical improvements, the in-orbit lifetime was also continuously increased. Key characteristics of each satellite type are summarized in Table 8.8. All GLONASS satellites were developed by the Academician Reshetnev

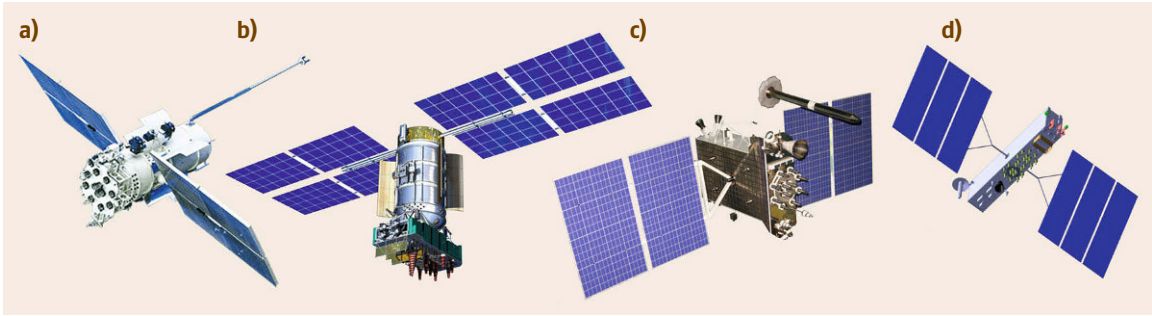


Fig. 8.15a–d The GLONASS satellites family: (a) *GLONASS IIV*, (b) *GLONASS-M*, (c) *GLONASS-K1* (d) *GLONASS-K2*. Artist's drawings (courtesy of ISS Reshetnev)

Research and Production Association of Applied Mechanics (NPO PM), which is now part of the joint stock company Information Satellite Systems (ISS)-Reshetnev Company.

While not part of the actual radionavigation system, it is worthwhile to note that the GLONASS constellation is complemented by two geodetic satellites named Etalon (measuring gauge). The ball-shaped satellites of 1.2 m diameter are completely passive and covered with corner cube reflectors for satellite laser ranging (SLR). The Etalon-1 and -2 satellites were launched in 1989 and injected into a typical GLONASS orbit along with two pairs of regular GLONASS navigation satellites. While initially used to study the orbital dynamics of satellites in medium altitude orbit, they still serve the international community today for fundamental studies of geodesy and dynamics of the Earth [8.44].

This section introduces the different spacecraft of the GLONASS family (Fig. 8.15) and describes their characteristic features and performances.

8.3.1 GLONASS I/II

The first generation of GLONASS navigation satellites (originally termed *Uragan*, the Russian word for hurricane) was developed in the late 1970s. Following [8.1], the GLONASS series comprises four subtypes, which are commonly designated as type (or block) I, IIa, IIb, and IIv (or IIc). A first type I satellite was launched in 1982 and the last IIv satellite was in use from 2005 to 2008.

The GLONASS spacecraft have a mass of roughly 1.4 t (including propellant for orbit maintenance) and are made up of a cylindrical structure with a length of about 3.3 m (Fig. 8.16). Two solar panels with a total surface area of about 24 m² delivered a net system power of about 1 kW. The GLONASS I/II satellites employed a pressurized platform design to protect the payload against the space environment. Heat dissipation was achieved through heat exchangers and four ther-

mal control flaps. The opening angle of these shutters could be varied and allowed the adjustment of the internal temperature with an accuracy of about 5 °C.

Attitude control was achieved through reaction wheels, which were periodically unloaded using magnetorquers. Reference measurements of the magnetic fields were provided by a magnetometer, which was mounted on an external boom (Fig. 8.16) to avoid magnetic disturbances by the satellite body. The GLONASS satellites were also equipped with a hydrazine propulsion system. It comprised two 5 N thrusters for orbit correction and 24 0.1 N thrusters for orientation changes and despinning after orbit injection. After reaching their assigned orbital slot, the satellites kept their nominal position within an argument-of-latitude deadband of $\pm 5^\circ$ throughout their operational lifetime with no need for further correction maneuvers.

The longitudinal $-x$ -axis of the cylindrical spacecraft body points towards the Earth and carries the L-band antenna. It is composed of 12 helix elements arranged in an inner ring with four elements and an outer

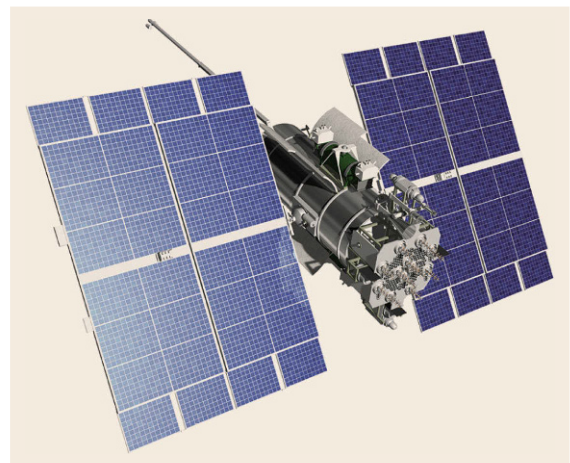


Fig. 8.16 Satellite of the first generation GLONASS system (courtesy of ISS Reshetnev)

Table 8.8 GLONASS satellites overview

Parameter	GLONASS	GLONASS-M	GLONASS-M+	GLONASS-K1	GLONASS-K1+	GLONASS-K2
First launch	1982	2003	2014	2012	2017	2016
Platform design	Pressurized	Pressurized	Pressurized	Unpressurized	Unpressurized	Unpressurized
Design lifetime (years)	3	7	7	10	10	> 10
Mass (kg)	1415	1415	1415+	995	995+	1645
System power (W)	1000	1450	1450	1600	1600	4370
Solar array size (m ²)	25	32	32	17	17	34
Pointing accuracy (Earth) (°)	0.5	0.5	0.5	0.5	0.5	0.25
Pointing accuracy (Sun) (°)	5	2	2	1	1	1
Navigation payload						
Mass (kg)	180	250	250	260	>260	520
Power consumption (W)	600	580	580	750	>750	2600
Clocks	(Rb), Cs	Cs	Cs	Cs, Rb	Cs, Rb	Cs, Rb
Clock stability (daily)	$5 \cdot 10^{-13}$	$1 \cdot 10^{-13}$	$1 \cdot 10^{-13}$	$(0.5 - 1) \cdot 10^{-13}$	$5 \cdot 10^{-14}$	$(0.5 - 5) \cdot 10^{-14}$
FDMA signals	L1, L2	L1, L2	L1, L2	L1, L2	L1, L2	L1, L2
CDMA signals	–	–	L3	L3	L2, L3	L1, L2, L3
Gross link	–	×	×	×	×	×
Laser reflector	×	×	×	×	×	×

ring with eight elements. These elements are phase coherently combined to achieve an M-shaped antenna gain pattern with a slightly higher beam intensity towards the rim of the Earth (Chap. 17). In addition, the GLONASS satellites carry a laser retroreflector array (LRA) for satellite laser ranging measurements [8.45]. The LRA comprises almost 400 individual corner cube reflectors accommodated between the L-band antenna elements to center the effective reflection point on the boresight axis.

In the GLONASS I series, two 5 MHz BERYL rubidium clocks with a stability of $5 \cdot 10^{-12}$ over one day served as the primary atomic frequency standard (AFS). The subsequent type II spacecraft were equipped with three GEM cesium clocks (offering a two-fold cold redundancy) that achieved a ten times better performance [8.46, 47]. The average lifetime of these early clocks amounted to only 1.5 years, which notably restricted the overall mission duration.

In total, 87 GLONASS I and II satellites were launched from 1982 to 2005. This includes six satellites that did not reach their final target orbit. Typical operations periods ranged from about one year for the early satellites to five years for the latest units. The first fully operational GLONASS constellation completed in 1995 was exclusively made up of IIv spacecraft.

8.3.2 GLONASS-M

From 2003 onwards, the first-generation GLONASS satellites were replaced by the modernized GLONASS-M series. These spacecraft share the same core structure and exhibit a similar mass as the GLONASS I/II series but are easily distinguished by the different placement of the solar panel rotation axis and a larger antenna panel on their front side (Fig. 8.17). Also, the satellites no longer carry a magnetometer boom on the zenith-facing side of the spacecraft body.

Like GLONASS I and II, the GLONASS-M satellites make use of a pressurized container and exhibit the characteristic thermal control flaps, but offer better mission and operational performances as well a longer design lifetime (seven years) than their predecessors. The GLONASS-M satellites achieve an 0.5° accuracy for the nadir pointing attitude control and a 2° pointing accuracy for the solar panels. The use of larger solar panels (34 m^2) provides a higher overall system power of 1.5 kW.

The laser retroreflector array uses a more compact design than on GLONASS I/II and is accommodated next to the L-band navigation antenna on the front panel with a small lateral offset from the principal body axis of the spacecraft (Fig. 8.18).

As a novel feature, GLONASS-M satellites include a radio-based intersatellite link [8.48], which is undergoing flight validation and will help to mitigate the limited geographical coverage of the GLONASS ground segment. It provides distance measurements between spacecraft based on dual one-way pseudoranges and can thus contribute to an improved ephemeris and clock accuracy [8.49].

The latest GLONASS-M satellites are also equipped with a prototype Inter-Satellite Laser Navigation and Communication System (ISLNCs, Fig. 8.18). Early in-flight experiments have demonstrated the capability to measure the distance between two satellites with a precision of about 3 cm and to synchronize the onboard clocks to each other with a corresponding error of less than 0.1 cm [8.50, 51].



Fig. 8.17 GLONASS-M spacecraft (courtesy of ISS Reshetnev)

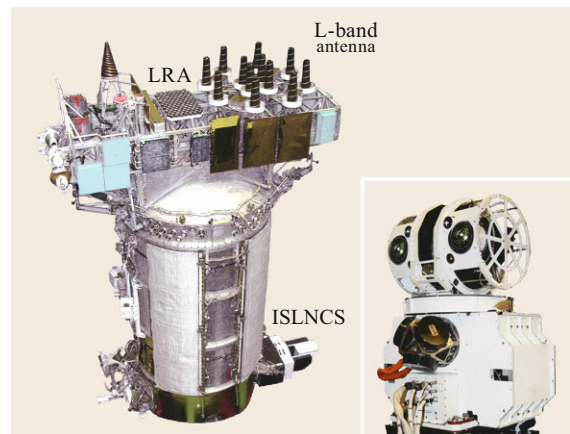


Fig. 8.18 GLONASS-M spacecraft with laser retroreflector array (LRA) and intersatellite laser navigation and communication system (ISLNCs). The insert shows the ISLNCs design in use since 2013 (courtesy of Science-Industry Corporation of Precise Device Engineering Systems (NPK SPP))

Besides the modernization of the spacecraft platform, the GLONASS-M satellites introduced various changes to the radio navigation signals [8.52]. The transmitted frequencies were shifted to a lower range $((1598.0625 \dots 1605.375) \pm 5.11 \text{ MHz}$ for L1 and $L2 = (1242.9375 \dots 11248.625) \pm 5.11 \text{ MHz}$ for L2) and filters were installed to reduce out-of-band emission in the frequency bands 1610.6–1613.8 MHz and 1660.0–1670.0 MHz down to the level provided in ITU recommendation 769 [8.31]. Furthermore, the transmission power in the L2 band was doubled and the civil navigation signal was added to L2, thus offering the first fully open dual-frequency navigation service. Along with these changes, various improvements to the navigation message were made. Available spare bytes in the original design were populated with new parameters such as the GPS-GLONASS time difference, leap second announcements or the age of orbit and clock data. Starting with spacecraft No. 55 in 2014, transmission of the L3OC was, furthermore, added to the GLONASS-M satellites.

Like their predecessors, the GLONASS-M satellites employ three cesium clocks as the primary frequency standard. According to the technical requirements the frequency stability amounts $1 \cdot 10^{-13}$ over one day, which directly contributes to a better overall navigation performance.

As of early 2015, the operational GLONASS constellation is entirely composed of GLONASS-M satellites.

8.3.3 GLONASS-K

The GLONASS-K series represents the latest generation of spacecraft in the GLONASS constellation. It comprises two subseries, namely the lighter K1 satellites (Fig. 8.19) and the more heavy K2 type with full capabilities. Two GLONASS-K1 satellites were launched in 2011 and 2014. In February 2016, the second GLONASS-K1 satellite was introduced into the constellation for nominal operation, while the first GLONASS-K1 satellite remains reserved for test purposes. Construction and deployment of the GLONASS-K2 satellites is planned for the second half of the decade.

GLONASS-K satellites are the first to use an unpressurized payload and service module. They build upon the Express-1000K spacecraft bus developed by ISS Reshetnev (formerly NPO PM) for various geostationary communication and relay satellites. The box-shaped structure is made up of lightweight hon-



Fig. 8.19 GLONASS-K1 spacecraft (courtesy of ISS Reshetnev)

eycomb panels and heat pipes are used for thermal control [8.48]. With a mass of 935 kg, the K1 satellites have only two thirds of the mass of their predecessors, which offers greater flexibility in the launcher selection. The use of advanced Ga-As solar cells enables a high electrical power (1.6 kW) despite a smaller size of the solar panels (17 m^2) than all previous satellites. The design lifetime of ten years is substantially longer than that of the earlier generations and will assist a smooth and interruption-free GLONASS service.

Similar to the GLONASS I/II spacecraft, the laser retroreflector of the GLONASS-K1 satellite is integrated into the L-band antenna structure to align both phase centers with the nadir-pointing axis through the center of mass. In this way, the phase center location is invariant under yaw rotations that are continuously performed to keep the solar panels perpendicular to the Sun (Sect. 3.4). The atomic frequency standards of the GLONASS-K1 satellites comprise two cesium and two rubidium clocks with a specified performance of $(0.5 - 1.0) \cdot 10^{-13}$. All GLONASS-K1 satellites transmit the CDMA L3 open service signal (L3OC) in addition to the legacy FDMA signals of the open and authorized services on L1 and L2. Support of L2 and, subsequently, L1 CDMA signals is foreseen for an enhanced version of K1+ satellites and the larger K2 satellites.

Aside from the core navigation payload, the GLONASS-K satellites are equipped with a radio cross link for data exchange and ranging, an optical cross link, a Cospas-Sarsat [8.53] distress alert detection and routing system, as well as an optical onboard system for simultaneous two-way and one-way laser ranging measurements intended for calibration and remote clock synchronization.

8.4 Launch Vehicles

The buildup and sustainment of the GLONASS constellation is provided through triple launches with the Proton launch vehicle or single launches with the Soyuz rocket (Fig. 8.20). Launches of the Soyuz rocket are performed from the Plesetsk Cosmodrome [8.54] in northern Russia (62.9° N, 40.6° E) and maintained by the Russian Air and Space Defence Forces. Proton launches, in contrast, are conducted from the rocket and space complex in Baikonur [8.55]. The latter site is located in the Republic of Kazakhstan (45.6° N, 63.3° E) and leased by the Russian Federation for its national space program.

The Proton-K rocket and its stronger and modernized M version are heavy-lift launch vehicles with a long flight heritage. The 3(+1)-stage rockets with a height of 53–58 m serve a wide range of missions in different orbits. They can carry up to 22 t into a low Earth orbit (LEO) or about 6 t into geostationary transfer orbit (GTO).

The first stage comprises six RD-275 engines with a common central oxidizer tank and six permanently attached strap-on fuel tanks. The stage provides a total thrust of 10 MN over a burn duration of 2 min and accelerates the rocket to ≈ 1.6 km/s before burnout at an altitude of about 40 km. The second stage combines four RD-0210/0211 engines with a total thrust of roughly 2 MN. It completes its firing approximately

5 min after liftoff near an altitude of 120 km. The third and final stage combines a single RD-0212 engine and a RD-0214 steering engine for fine control of the injection vector and velocity. All of the aforementioned engines use unsymmetrical dimethyl hydrazine (UDMH, $C_2H_8N_2$) and nitrogen tetroxide (N_2O_4) as fuel and oxidizer, which are known to be highly toxic but do not require cooling for storage.

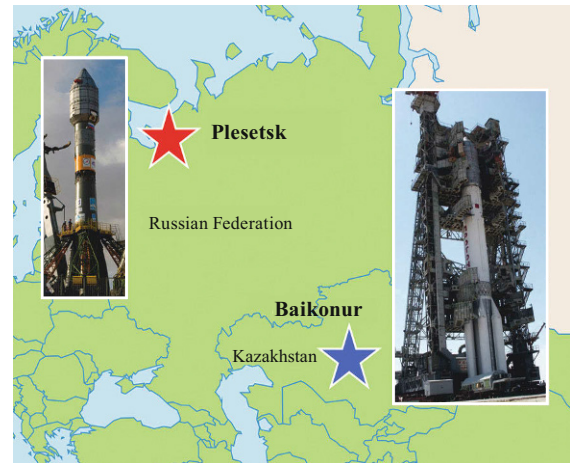


Fig. 8.20 Launch sites and vehicles for the GLONASS navigation satellites (courtesy of ISS Reshetnev)

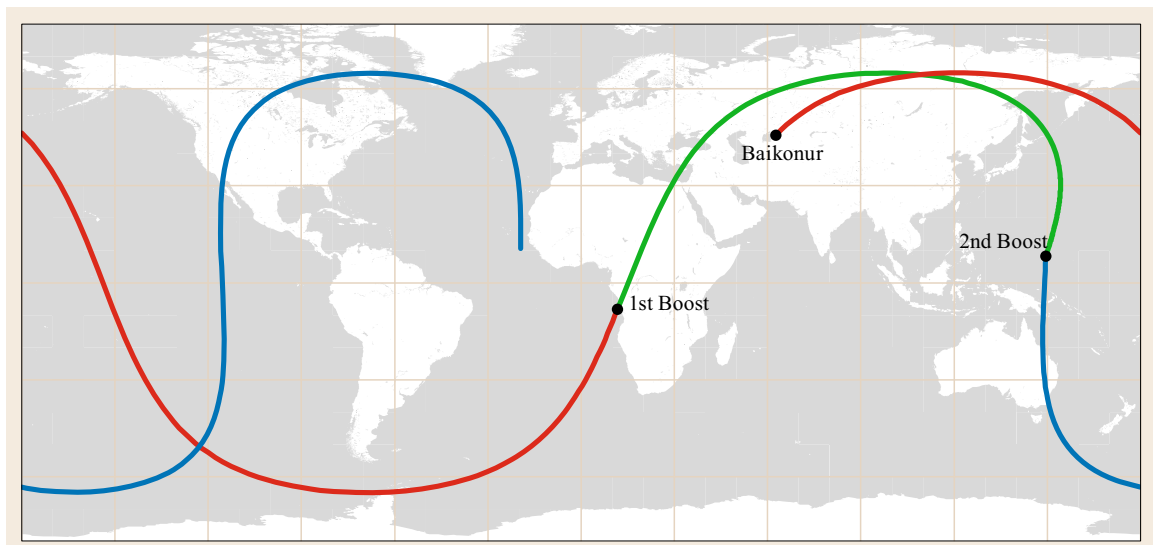


Fig. 8.21 Representative example of the ground track during orbit injection of GLONASS satellites with a Proton rocket. The low Earth parking orbit and the elliptical transfer orbit are marked by red and green lines respectively, with gray dots indicating the approximate location of the upper stage burns. The blue line describes the first orbital revolution of the GLONASS spacecraft after separation. The illustration is based on patched Keplerian orbits and does not take into account the actual boost durations



Fig. 8.22 Integration of three GLONASS-M satellites with a Block-DM upper stage (courtesy of ISS Reshetnev)

Aside from the three main stages, the Proton rocket typically employs a reignitable upper stage to realize more complex mission profiles. For GLONASS launches, either the *DM* or *Breeze-M* upper stage are employed, which offer a thrust level of 80 and 20 kN respectively. A typical missions scenario for GLONASS orbit injection is illustrated in Fig. 8.21 based on [8.1, 56]. After ascent of the launcher and burnout of the third stage (roughly 10 min after launch), the upper stage with the attached GLONASS satellites is separated and moves around the Earth at an altitude of about 200 km. This parking orbit has an inclination of $\approx 64.8^\circ$ in accord with the orbit planes of the GLONASS constellation. A first boost of the upper stage raises the apogee to the desired target altitude of 19 130 km. The highly elliptical transfer orbit has an eccentricity of $e \approx 0.6$ and it takes about 3 h to proceed from perigee to apogee. Here, a second burn is performed, which circularizes the orbit. Thereafter, the GLONASS satellites are deployed into their target orbit. Following the separation from the upper stage, the solar panel deployment and the checkout of all onboard systems, the satellites are ultimately moved to their desired position in the orbital plane in a series of small

orbit correction maneuvers using their own thruster system.

The large payload capacity of the Proton rocket enables triple launches of GLONASS I/II or GLONASS-M satellites, which was particularly helpful during the buildup and rebuild phase of the constellation. To fit three satellites under the Proton fairing, the solar panels are folded around the satellites in a rhombic shape that enables a tight packing of the triplet on the launch adapter (Fig. 8.22). Out of 51 launches with a total of 132 GLONASS satellites conducted up to 2015, a wide majority were conducted with Proton-K and -M rockets carrying three GLONASS satellites (with occasional replacements by mass dummies or Etalon satellites) at a time.

Starting with the launch of GLONASS-K1 in 2011, the Soyuz-2b [8.54] was introduced as an alternative launch vehicle for GLONASS satellites. The rocket has a height of about 46 m and can carry a payload of 4–8 t into low Earth orbit (depending on the inclination and actual altitude). The Soyuz rocket carries four RD-107A strap-on boosters that serve as the first stage and develop a thrust of 3.3 MN over a 2 min boost phase. The central second and third stages are based on a single RD-108 and RD-0124 engine respectively, which provide thrust levels of 0.9 and 0.3 MN. Unlike the Proton rocket, the Soyuz launcher employs kerosene and liquid oxygen as propellant. A Fregat upper stage serves as forth stage and performs the transfer orbit injection as well as the final circularization of the orbit at the target altitude of the GLONASS constellation.

The Soyuz vehicle can carry a single GLONASS-M or -K1 satellite and has typically been used for replacing individual aged satellites in the constellation. As a side note, the Soyuz launcher has also been used for in-orbit delivery of various satellites of the Galileo constellation. Aside from the two precursor satellites Galileo in orbit validation element GIOVE-A and -B, most of these launches were, however, conducted from the European spaceport at Kourou, French Guiana.

8.5 Ground Segment

The ground segment is an essential part of the GLONASS architecture providing system operation and ultimate GLONASS performance. There is no formal division between system control and mission control functions. All operational procedures after satellite liftoff are fulfilled by the Air and Space Defense Forces (ASDF). If satellites are launched from the Baikonur launch site, the initial active phase of the launcher trajectory tracking is supported by the Russian Federal Space Agency (RFSa) assets.

The main functions of the GLONASS ground segment comprise:

- Support of operations in the launch and early orbit phase (LEOP)
- Commissioning of satellites and their transfer to dedicated orbit slots (in case of a triple launch or commissioning at an orbital spare position)
- Telemetry monitoring, command and control
- Mission planning and constellation keeping

- Maintenance and decommissioning of satellites at their end of life
- Monitoring of the ground assets status
- Generation of the system timescale and its steering to UTC(SU)
- Generation of orbit and clock data
- Upload of navigation data to the satellites
- Improvement of the satellite dynamics models
- Monitoring of the GLONASS navigation, positioning and timing performance
- Serving of external interfaces with civil institutions.

Core components of the GLONASS ground segment include the GLONASS system control center (SCC) and the central clocks (CCs), the telemetry, tracking and command stations (TT and C) and the uplink stations, as well as one-way monitoring stations and satellite laser ranging (SLR) stations. All major ground segment assets are located within the Russian territory at the ASDF sites shown in Fig. 8.23.

The system control center is located in Krasnoznamensk (formerly known as the closed town of Golitsyno-2) some 40 km southwest of the Moscow city center. It performs the planning and coordinates the work of all ground segment elements. Orbit determination and clock synchronization is implemented through processing of all available sets of data including two-way ranging from TT and C and uplink

stations, downloaded radio cross-link ranging data, and one-way ranging data from ASDF monitoring stations. Use of data from the RFSA monitoring stations is planned to improve orbit and clock information as well as the cross-link ranging data.

GLONASS System Time is maintained by the central clock facility. It includes a group of four hydrogen masers, which are continuously steered to the Russian realization of Coordinated Universal time, UTC(SU) [8.27,28]. The master central clock at Shelkovo near Moscow is complemented by a second facility at Komsomolsk in the far east of Russia.

The telemetry, tracking and control stations are used to receive status information from the GLONASS satellites, to send control commands, and to perform two-way ranging measurements for orbit determination. For maximum coverage, a total of five TT and C stations situated in the western, central and eastern parts of Russia are available. They are complemented by five uplink stations in Shelkovo, Yeniseysk, Komsomolsk, Vorkuta, and Petropavlovsk. Each of these sites is equipped with two antennas, which enables up to three uploads of orbit and clock data to each GLONASS satellite per day.

The monitoring stations collect one-way pseudorange and carrier-phase measurements for orbit and clock determination as well as offline performance and integrity monitoring. Most of them are colo-

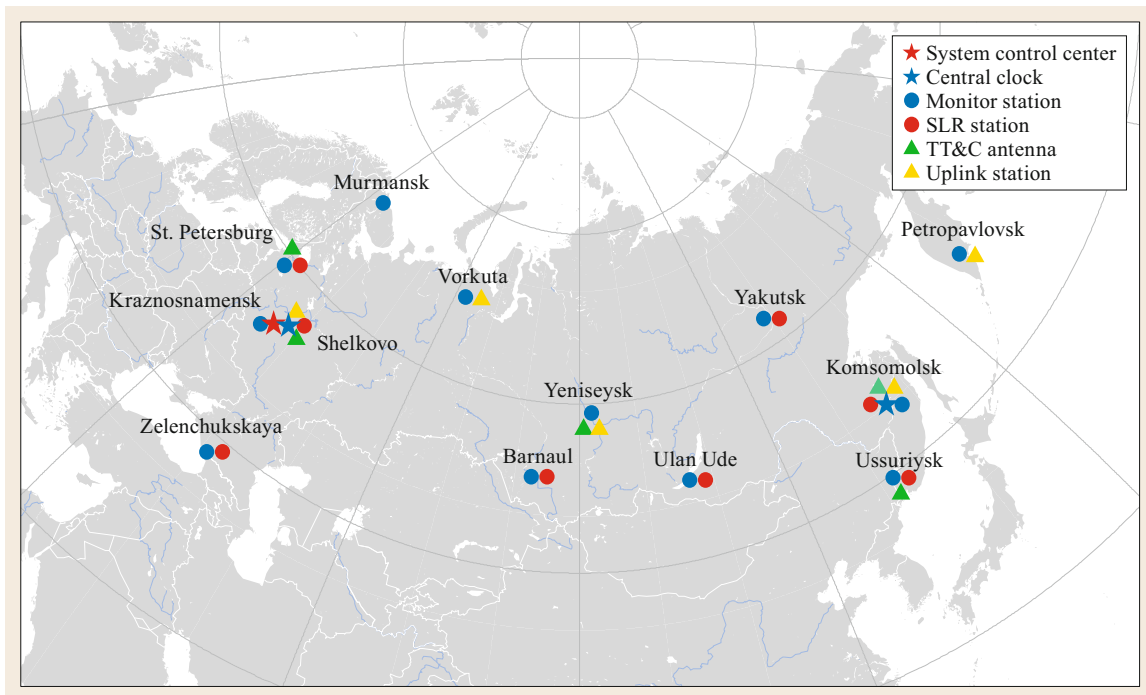


Fig. 8.23 GLONASS ground segment sites

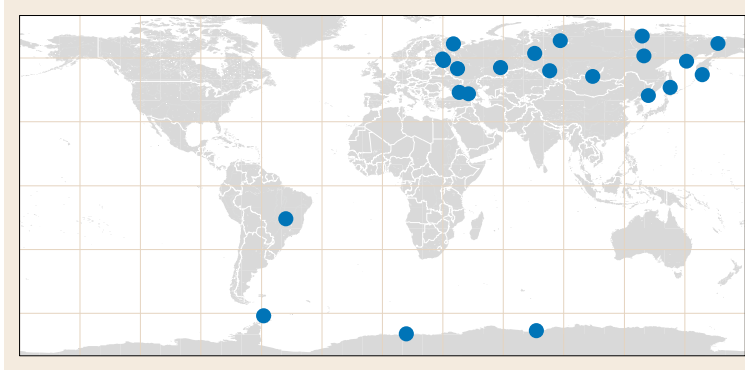


Fig. 8.25 Network of SDCM monitoring stations (status end-2014)



Fig. 8.24 Altay Laser Ranging Center near Barnaul (Altay) (courtesy of Science-Industry Corporation of Precise Device Engineering Systems (NPK SPP))

cated with laser ranging stations (Fig. 8.24), which enables complementary optical two-way distance measurements. The SLR observations are used for calibration of radiometric distance measurements as well as orbit determination and accuracy validation. They also contribute to an improved realization of the GLONASS reference frame. The routine use of SLR for GLONASS operations [8.57] is unique among all navigation satellite systems, but forms an integral part of the system architecture. From the very beginning, all GLONASS satellites were equipped with laser retroreflectors and the high-accuracy SLR measurements have helped in coping with the limited geographical distribution and accuracy of conventional radiometric tracking stations. In total, 13 monitoring and nine laser ranging stations distributed over the Russian territory (and neighboring states of the former Soviet Union) are currently included in the GLONASS ground segment.

As part of the ongoing GLONASS modernization and enhancement, the system for differential correction and monitoring (SDCM) [8.58] has been established. SDCM is based on a network of reference stations, which are equipped with combined GPS/GLONASS dual-frequency receivers, hydrogen maser atomic clocks and direct communication links for real-time data transfer. By the end of 2014, a total of 18 stations have been deployed in Russian territory as well as four stations in Antarctica and Brazil (Fig. 8.25). Further stations are planned in Cuba and Kazakhstan as well as other countries in South America, Africa and Asia/Oceania to achieve a worldwide coverage. The SDCM network enables a continuous performance and integrity monitoring [8.59] as well as real-time corrections for precise point positioning applications [8.60]. SDCM correction data are provided via the Internet for terrestrial users and through the Luch-5A/B relay and communication satellites. Incorporation of the SDCM reference stations into the GLONASS orbit and clock determination could constitute an important building block for a fully global, high-performance GLONASS navigation service.

For synchronization of the onboard clocks, support of one-way laser ranging is currently under development [8.51]. Using photodetectors on board the GLONASS satellites, the arrival time of laser pulses with precisely known transmission time can be measured relative to the onboard timescale. By comparing these measurements with traditional two-way SLR observations, the difference between the satellite and ground clocks can then be determined [8.61]. Along with the installation of laser time transfer equipment in the future GLONASS satellites, the GLONASS ground segment will be upgraded to support such operations on a routine basis.

8.6 GLONASS Open Service Performance

Continuous monitoring of the GLONASS system performance and integrity is presently provided through various institutions and services in Russia. These include the Information and Analysis Centre of the Russian Federal Space Agency (IAC [8.62]), the System for Differential Correction and Monitoring (SDCM [8.59]), and the System for High-accuracy Determination of Ephemeris and Time Corrections (SVOEVP [8.63, 64]). Complementary to Russian reference stations, these monitoring services make use of stations from the International GNSS Service (IGS [8.65]) network to enable a fully global coverage. Unless otherwise noted, the performance results presented in this section are based on analyses and data of the IAC.

In a statistical sense, the positioning performance of a satellite navigation system can be expressed as the product of the position dilution of precision (PDOP) and the user range error (URE). The PDOP depends only on the number of tracked GNSSs and the geometric distribution of their line-of-sight vectors. The URE in contrast describes the root mean square errors of the difference between modeled and observed pseudoranges. Aside from user equipment errors

(UEE) such as noise and multipath or uncompensated atmospheric delays, it comprises the signal-in-space range error (SISRE). The latter describes the impact of errors in the broadcast orbit and clock parameters on the range computation.

Since completion of the nominal 24-satellite constellation in 2012, GLONASS provides global daily availability of a better than 99% at a 5° elevation mask angle and a PDOP of less than 6. An example of the instantaneous PDOP map for the current GLONASS constellation is shown in Fig. 8.26. PDOP values amount to 1.5–2.5 for most of the globe and only exceed a value of 3 in very limited geographic regions.

The signal-in-space range error as monitored by the IAC exhibits typical variations in the range of 1–2 m across individual satellites of the constellation (Fig. 8.27). For comparison, a mean SISRE of 1.9 m has been derived from the analysis of broadcast ephemeris data covering a one year interval in 2013/2014 [8.66] while SISRE values of 1–4 m were obtained by [8.67] for individual GLONASS satellites in the 2009–2011 time frame.

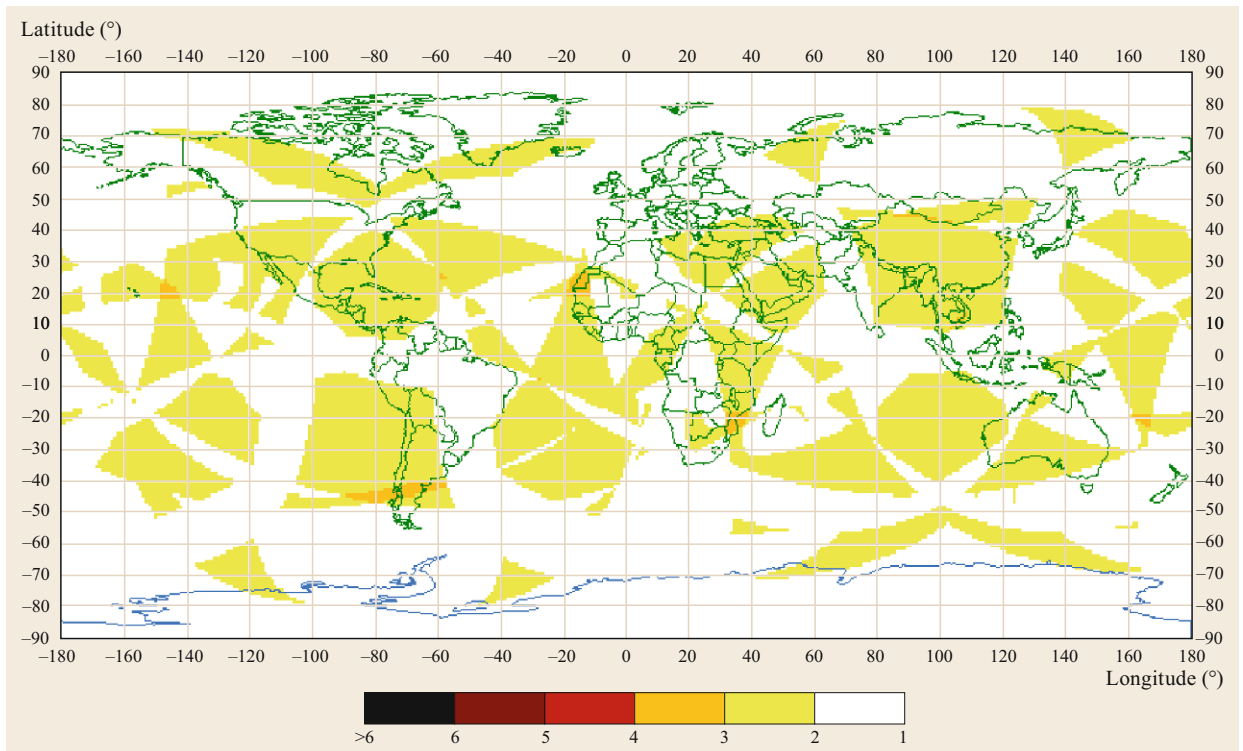


Fig. 8.26 Instantaneous PDOP map for the GLONASS constellation on 8 March 2015, 10:30 UTC (mask elevation angle 5°) (courtesy of Federal Space Agency-Information Analytical Centre)

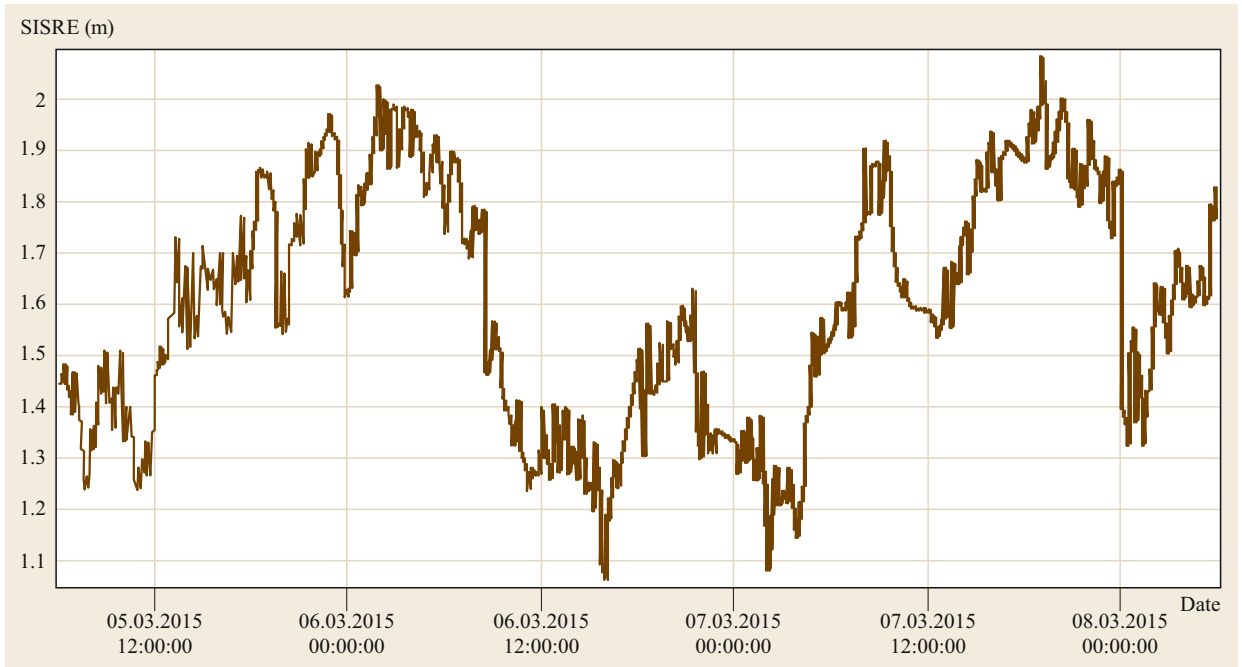


Fig. 8.27 Average GLONASS constellation SISRE for 6–8 March 2015 (RMS, m) (courtesy of Federal Space Agency-Information Analytical Centre)

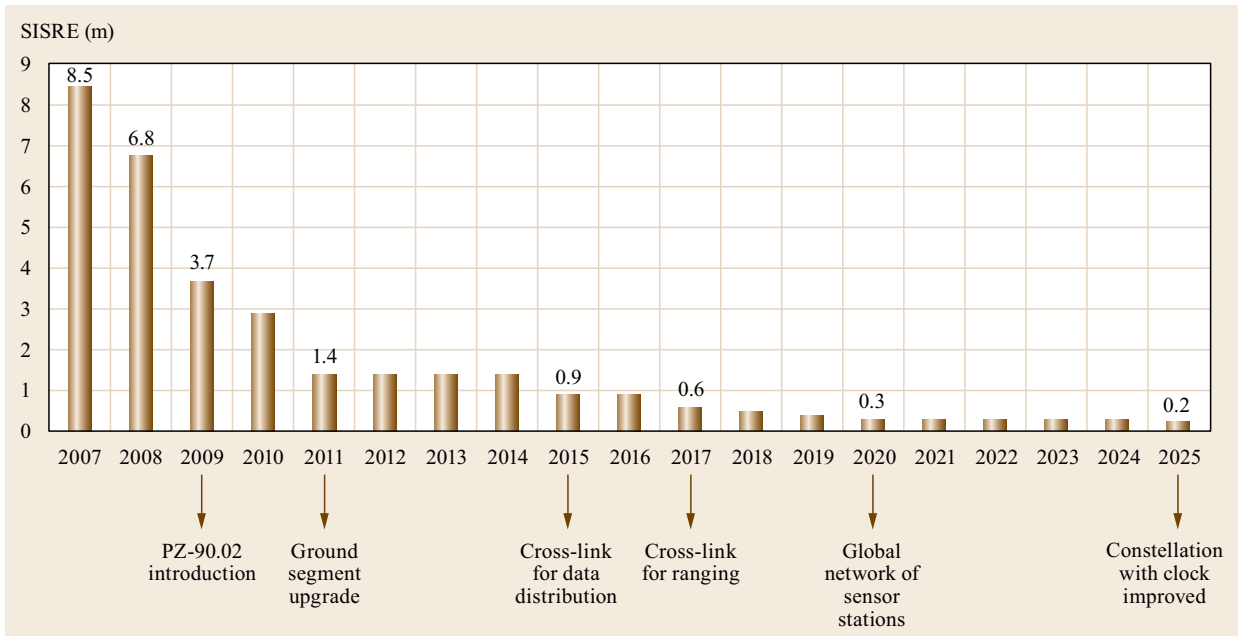


Fig. 8.28 Mean SISRE of the GLONASS constellation through GPI plan implementation

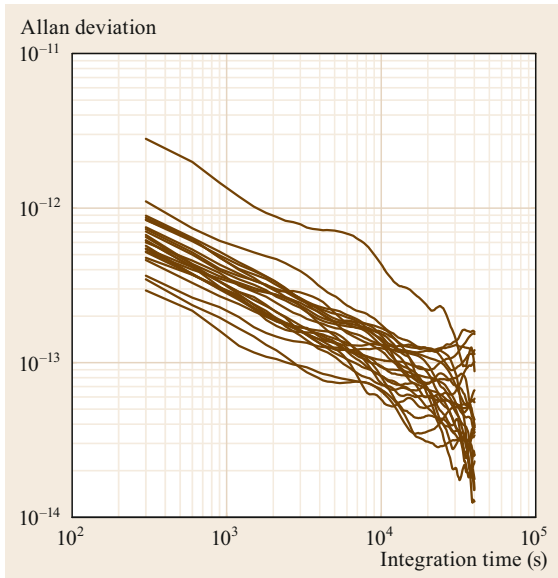


Fig. 8.29 Allan deviation of GLONASS-M satellites on 5 March 2015 as derived from European Space Agency (ESA) clock products (courtesy of P. Steigenberger (DLR))

One of the key factors contributing to the SISRE is the onboard clock stability. It is commonly characterized by the Allan deviation (ADEV),

that is the relative frequency error over a specified time interval (Fig. 8.29). For the current constellation of GLONASS-M satellites, the Allan deviation (ADEV) over a one-day correlation time is typically better than $(0.5 - 1.0) \cdot 10^{-13}$ [8.62, 68]. Over short timescales of 1–100 s, ADEV values of about $1 \cdot 10^{-11}$ have been reported in [8.69, 70]. An overall improvement of the onboard clock stability is expected from new types of atomic frequency standards on the next-generation GLONASS-K satellites.

The GLONASS performance improvement (GPI) plan foresees a continuous SISRE reduction to less than 0.5 m by 2020 (Fig. 8.28). Key steps to achieve this improvement include the processing of carrier-phase measurements in the orbit determination and time synchronization process, the use of intersatellite communication cross links for performing up to 24 navigation data uploads per day, use of intersatellite ranging data for orbit and clock determination in the ground segment facilities, and the global expansion of the monitoring network. Further improvements are expected from the introduction of high-performance CDMA signals and new generations of onboard clocks.

References

- 8.1 N.L. Johnson: GLONASS spacecraft, *GPS World* **5**(11), 51–58 (1994)
- 8.2 V.V. Dvorkin, Y.I. Nosenko, Y.M. Urlichich, A.M. Finkel'shtein: The Russian global navigation satellite program, *Her. Russ. Acad. Sci.* **79**(1), 7–13 (2009)
- 8.3 T.G. Anodina: *The GLONASS System Technical Characteristics and Performance* (International Civil Aviation Organization, Montreal, Canada 1988), Working Paper FANS/4-WP/75
- 8.4 S.A. Dale, P. Daly: The Soviet Union's GLONASS navigation satellites, *IEEE Aerosp. Electron. Syst. Mag.* **2**(5), 13–17 (1987)
- 8.5 G.R. Lennen: The USSR's GLONASS P-code-determination and initial results, *ION GPS 1989*, Colorado Springs (ION, Virginia 1989) pp. 77–83
- 8.6 S.A. Dale, P. Daly, I.D. Kitching: Understanding signals from GLONASS navigation satellites, *Int. J. Sat. Commun.* **7**(1), 11–22 (1989)
- 8.7 Global Navigation Satellite System GLONASS – Interface Control Document, v5.1, (Russian Institute of Space Device Engineering, Moscow, 2008)
- 8.8 Y. Urlichich, V. Subbotin, G. Stupak, V. Dvorkin, A. Povaliaev, S. Karutin: GLONASS modernization, *ION GNSS 2011*, Portland (ION, Virginia 2010) pp. 3125–3128
- 8.9 V. Putin: On Use of GLONASS (Global Navigation Satellite System) for the Benefit of Social and Economic Development of the Russian Federation, Presidential Decree No. 638, Kremlin, Moscow (2007)
- 8.10 T. Mirgorodskaya: GLONASS and critical infrastructure, *Proc. 9th Meet. Int. Comm. GNSS (ICG)*, Work. Group A, Prague (UNOOSA, Vienna 2014)
- 8.11 N. Zarraoa, W. Mai, E. Sardon, A. Jungstand: Preliminary evaluation of the Russian GLONASS system as a potential geodetic tool, *J. Geod.* **72**(6), 356–363 (1998)
- 8.12 P. Willis, J. Slater, G. Beutler, W. Gurtner, C. Noll, R. Weber, R.E. Neilan, G. Hein: The IGEX-98-campaign: Highlights and perspective. In: *Geodesy Beyond 2000, International Association of Geodesy Symposia*, Vol. 121, ed. by K.-P. Schwarz (Springer, Berlin 2000) pp. 22–25
- 8.13 R. Weber, J.A. Slater, E. Fagnier, V. Glotov, H. Habrich, I. Romero, S. Schaer: Precise GLONASS orbit determination within the IGS/IGLOS-pilot project, *Adv. Space Res.* **36**(3), 369–375 (2005)
- 8.14 J.G. Walker: Satellite constellations, *J. Br. Interplanet. Soc.* **37**, 559–572 (1984)
- 8.15 Parametry Zemli 1990 goda. Version PZ-90.11 (Earth Model PZ-90.11; In Russian). Military Topography Agency of the General Staff of the Armed Forces of the Russian Federation (Moscow 2014) <http://structure.mil.ru/files/pz-90.pdf>
- 8.16 S. Fearheller, J. Purvis, R. Clark: The Russian GLONASS system. In: *Understanding GPS – Principles and Applications*, ed. by E.D. Kaplan (Artech House,

- Boston, London 1996) pp. 439–465
- 8.17 V. Vdovin, A. Dorofeeva: Global geocentric coordinate system of the Russian federation, Proc. 7th Meet. Int. Comm. GNSS (ICG), Work. Group D, Beijing (UNOOSA, Vienna 2012)
 - 8.18 A.N. Zueva, E.V. Novikov, D.I. Pleshakov, I.V. Gusev: System of geodetic parameters parametry zemli 1990 PZ–90.11, Proc. 9th Meet. Int. Comm. GNSS (ICG), Work. Group D, Prague (UNOOSA, Vienna 2014)
 - 8.19 P.N. Misra, R.I. Abbot, E.M. Gaposcbkin: Integrated Use of GPS and GLONASS: Transformation between WGS 84 and PZ–90, ION GPS 1996, Kansas City (ION, Virginia 1996) pp. 307–314
 - 8.20 U. Rossbach, H. Habrich, N. Zarraoa: Transformation Parameters between PZ–90 and WGS 84, ION GPS 1996, Kansas City (ION, Virginia 1996) pp. 279–285
 - 8.21 C. Boucher, Z. Altamimi: ITRS, PZ–90 and WGS 84: Current realizations and the related transformation parameters, *J. Geod.* **75**(11), 613–619 (2001)
 - 8.22 S.G. Revniyykh: GLONASS status and progress, Proc. 47th CGSIC Meet., Fort Worth (CGSIC, Alexandria 2007)
 - 8.23 Global Navigation Satellite System and Global Positioning System: Coordinate Systems, Methods of Transformations for Determinated Points Coordinate; STB GOST Standard 51794–2008 (Federalnoje agentstwo po technitscheskomu regulirowaniju i metrologii, Moscow, 2008) in Russian
 - 8.24 Yu. Domnin, B. Gaigerov, N. Koshelyaevsky, S. Poushkin, F. Rusin, V. Tatarenkov, G. Yolkin: Fifty years of atomic time-keeping at VNIIFTRI, *Metrologia* **42**(3), S55–S63 (2005)
 - 8.25 I. Blinov, Y. Domnin, S. Donchenko, N. Koshelyaevsky, V. Kostromin: Progress at the state time and frequency standard of Russia, European Frequency and Time Forum (EFTF) 2012, Gothenburg (2012) pp. 144–147
 - 8.26 W. Lewandowski, E.F. Arias: GNSS times and UTC, *Metrologia* **48**(4), S219–S224 (2011)
 - 8.27 A. Shchipunov: Generating and transferring the national time scale in GLONASS, ION GNSS 2012, Nashville (ION, Virginia 2012) pp. 3950–3962
 - 8.28 A.V. Druzhin, V. Palchikov: Current state and perspectives of UTC(SU) broadcast by GLONASS, Proc. 9th Meet. Int. Comm. GNSS (ICG), Prague (UNOOSA, Vienna 2014) pp. 1–9
 - 8.29 A. Bolkonov: GLONASS open service performance parameters standard and GNSS open service performance parameters template status, Proc. 9th Meet. Int. Comm. GNSS (ICG), Work. Group A, Prague (UNOOSA, Vienna 2014)
 - 8.30 R.B. Langley: GLONASS: Review and update, *GPS World* **8**(11), 51–58 (1994)
 - 8.31 Protection criteria used for radio astronomical observations, Recommendation RA 769, rev. 2, May 2003 (ITU, 2003) <http://www.itu.int/rec/R-REC-RA.769/en/>
 - 8.32 J. Galt: Interference with Astronomical Observations of OH Masers from the Soviet Union's GLONASS satellites. In: *IAU Colloq. 112 Light Pollution, Radio Interference, and Space Debris*, ed. by D.L. Crawford (IAU, Paris 1991) pp. 213–221
 - 8.33 J.A. Ávila Rodríguez: On Generalized Signal Waveforms for Satellite Navigation, Ph.D. Thesis (Univ. der Bundeswehr, Neubiberg 2008)
 - 8.34 B.A. Stein: PRN codes for GPS/GLONASS: A comparison, ION NTM 1990, San Diego (ION, Virginia 1990) pp. 31–35
 - 8.35 J. Beser, J. Danaher: The 3S navigation R-100 family of integrated GPS/GLONASS receivers: Description and performance results, ION NTM 1993, San Francisco (ION, Virginia 1993) pp. 25–45
 - 8.36 P. Daly, S. Riley: GLONASS P-code data message, ION NTM 1994, San Diego (ION, Virginia 1994) pp. 195–202
 - 8.37 S. Zaminpardaz, P.J.G. Teunissen, N. Nadarajah: GLONASS CDMA L3 ambiguity resolution and positioning, *GPS Solut.* (2016) doi:[10.1007/s10291-016-0544-y](https://doi.org/10.1007/s10291-016-0544-y)
 - 8.38 Y. Urlichich, V. Subbotin, G. Stupak, V. Dvorkin, A. Povaliaev, S. Karutin: GLONASS developing strategy, ION GNSS 2010, Portland (ION, Virginia 2010) pp. 1566–1571
 - 8.39 S. Karutin: GLONASS Signals and Augmentations, ION GNSS 2012, Nashville (ION, Virginia 2012) pp. 3878–3911
 - 8.40 T. Kasami: Weight Distribution Formula for Some Class of Cyclic Codes, Tech. Rep. R285 (Univ. Illinois, Illinois 1966) pp. 1–24
 - 8.41 T. Hellesteth, P.V. Kumar: Pseudonoise sequences. In: *The Mobile Communications Handbook*, ed. by J.D. Gibson (CRC, Boca Raton 1999) pp. 237–252
 - 8.42 S. Thoelet, S. Erker, J. Furthner, M. Meurer, G.X. Gao, L. Heng, T. Walter, P. Enge: First signal in space analysis of GLONASS K-1, ION GNSS 2011, Portland (ION, Virginia 2011) pp. 3076–3082
 - 8.43 A.A. Povalyaev: GLONASS navigation message format for flexible row structure, ION GNSS 2013, Nashville (ION, Virginia 2013) pp. 972–974
 - 8.44 G.M. Appleby: Orbit determinations of the lageos and etalon satellites – A comparison of geodetic results and orbital evolution of the etalons, dynamics and astrometry of natural and artificial celestial bodies, Proc. Conf. Astrom. Celest. Mech., Poznan 1993, ed. by K. Kurzynska, F. Barlier, P.K. Seidelmann, I. Wyrtrzyaszczak (IAU, Pairs 1994)
 - 8.45 T. Otsubo, G.M. Appleby, P. Gibbs: GLONASS laser ranging accuracy with satellite signature effect, *Surv. Geophys.* **22**(5/6), 509–516 (2001)
 - 8.46 Y.G. Gouzhva, A.G. Gevorkyan, P.P. Bogdanov: Accuracy estimation of GLONASS satellite oscillators, Proc. 46th Freq. Control Symp., Hershey (1992) pp. 306–309
 - 8.47 A.B. Bassevich, P.P. Bogdanov, A.G. Gevorkyan, A.E. Tyulyakov: GLONASS onboard time/frequency standards: Ten years of operation, Proc. 28th Ann. PTI Meet., Reston (DTIC, Fort Belvoir 1996) pp. 455–462
 - 8.48 R. Fatkulin, V. Kossenko, S. Storozhev, V. Zvonar, V. Chebotarev: GLONASS space segment: Satellite constellation, GLONASS–M and GLONASS–K spacecraft, main features, ION GNSS 2012, Nashville (ION, Virginia 2012) pp. 3912–3930
 - 8.49 A. Bolkunov, I. Zolkin, E. Ignatovich, A. Schekutiev: Intersatellite links as critical element of advanced satellite navigation technologies, *Sci. Tech. J. 'Polyot' (Flight)* **4**, 29–33 (2013)
 - 8.50 A. Chubykin, S. Dmitriev, V. Shargorodskiy, V. Sumerin: Intersatellite laser navigating link system, Proc. WPLTN Tech. Workshop One-Way Two-Way SLR GNSS Co-located RF Tech., St.Petersburg

- (2012) pp. 1–18
- 8.51 V.D. Shargorodsky, V.V. Pasynkov, M.A. Sadovnikov, A.A. Chubykin: Laser GLONASS: Era of extended precision, *GLONASS Herald* **14**, 22–26 (2013)
 - 8.52 G.M. Polischuk, V.I. Kozlov, V.V. Ilitchov, A.G. Kozlov, V.A. Bartenev, V.E. Kossenko, N.A. Anphimov, S.G. Revnivikh, S.B. Pisarev, A.E. Tyulyakov: The global navigation satellite system GLONASS: Development and usage in the 21st century, *Proc. 34th PTI Meet.* 2002, Reston (DTIC, Fort Belvoir 2002) pp. 39–50
 - 8.53 D.S. Ilcev: Cospas–Sarsat LEO and GEO: Satellite distress and safety systems (SDSS), *Int. J. Satell. Commun. Netw.* **25**(6), 559–573 (2007)
 - 8.54 Th. Pirard: Space centres–launch sites: The USSR. In: *The Cambridge Encyclopedia of Space*, ed. by M. Rycroft (Cambridge Univ. Press, Cambridge 1990) pp. 126–127
 - 8.55 Y. Tchourianov: *Baikonur – The Advent of a New Century* (Voennyi parad, Moscow 2005)
 - 8.56 S. Revnivikh: GLONASS status and progress, *Proc. CGSIC Meet.*, Savannah (2008)
 - 8.57 V. Burmistrov, A. Fedotov, N. Parkhomenko, V. Pasinkov, V. Shargorodsky, V. Vasiliev: The Russian laser tracking network, *Proc. 15th ILRS Workshop 2006*, Canberra (2006) pp. 1–3
 - 8.58 G. Stupak: SDCM status and plans, *Proc. 7th Meet. Int. Comm. GNSS (ICG)*, Beijing (UNOOSA, Vienna 2012) pp. 1–15
 - 8.59 Russian System of Differential Correction and Monitoring (SDCM): http://www.sdc.ru/index_eng.html
 - 8.60 V.V. Dvorkin, S.N. Karutin: Construction of a system for precise determination of the position of users of global navigation satellite systems, *Meas. Tech.* **54**(5), 517–523 (2011)
 - 8.61 M.A. Sadovnikov, V.D. Shargorodskiy: Stages of development of stations, networks and SLR usage methods for global space geodetic and navigation systems in Russia, *Proc 19th ILRS Workshop 2014*, Annapolis (2014) pp. 1–23
 - 8.62 Positioning, Navigation and Timing Information and Analysis Centre, GLONASS system status information: <http://www.glonass-center.ru/en/>
 - 8.63 A.Y. Suslov, E.V. Titov, A.A. Fedotov, V.D. Shargorodskiy: System for high-accuracy determination of ephemeris and time corrections (SVOEVP) GLONASS, *Proc. WPLTN Tech. Workshop One-Way Two-Way SLR GNSS Co-located RF Tech.*, St. Petersburg (2012) pp. 1–18
 - 8.64 GLONASS navigation performance information: <http://www.glonass-svoevp.ru/Func/plotnostil/>
 - 8.65 J.M. Dow, R.E. Neilan, C. Rizos: The International GNSS Service in a changing landscape of global navigation satellite systems, *J. Geod.* **83**(3/4), 191–198 (2009)
 - 8.66 O. Montenbruck, P. Steigenberger, A. Hauschild: Broadcast versus precise ephemerides: A multi-GNSS perspective, *GPS Solutions* **19**(2), 321–333 (2015)
 - 8.67 L. Heng, G.X. Gao, T. Walter, P. Enge: Statistical characterization of GLONASS broadcast clock errors and signal-in-space errors, *ION ITM 2012*, Newport Beach (ION, Virginia 2012) pp. 1697–1707
 - 8.68 M. Fritsche, K. Sośnica, C.J. Rodríguez-Solano, P. Steigenberger, K. Wang, R. Dietrich, R. Dach, U. Hugentobler, M. Rothacher: Homogeneous reprocessing of GPS, GLONASS and SLR observations, *J. Geod.* **88**(7), 625–642 (2014)
 - 8.69 A. Hauschild, O. Montenbruck, P. Steigenberger: Short-term analysis of GNSS clocks, *GPS Solutions* **17**(3), 295–307 (2013)
 - 8.70 E. Griggs, E.R. Kursinski, D. Akos: Short-term GNSS satellite clock stability, *Radio Sci.* **50**(8), 813–826 (2015)

Galileo

9. Galileo

Marco Falcone, Jörg Hahn, Thomas Burger

The European global navigation satellite system Galileo is designed as a self-standing satellite-based positioning system for worldwide service. It is independent from other systems with respect to satellite constellation, ground segment, and operation. Galileo is prepared to be compatible and interoperable with other radio navigation satellite systems, with global positioning system (GPS) as the main example. It uses the same physical principles as GPS, GLONASS, and others, that is radio signal-based ranging measurements from high-precision clocks as sources in orbit. The features of the first generation of Galileo comprise technological advances such as passive maser clock technology in orbit, plus modern system and signal concepts aligned to the planned and ongoing modernization of other systems. To the user, Galileo provides navigation signals on three frequencies E1, E6, and E5. The signals in E1 and E5 are coordinated with GPS L1 and L5, and both systems use equivalent modulation principles. This is expected to result in a benefit with respect to po-

9.1	Constellation	248
9.2	Signals and Services	250
9.2.1	Signal Components and Modulations.....	251
9.2.2	Navigation Message and Services	256
9.2.3	Ranging Performance	258
9.2.4	Timing Accuracy	263
9.3	Spacecraft	265
9.3.1	Satellite Platform	266
9.3.2	Satellite Payload Description	266
9.3.3	Launch Vehicles	268
9.4	Ground Segment	269
9.5	Summary	270
	References	271

sitioning accuracy, and in increased robustness of a positioning service derived from the combined use of multiple independent radio navigation systems. This chapter describes architecture and operations of Galileo.

The enormous potential benefits of satellite navigation for the citizens brought the European Space Agency (ESA) and the European Commission (EC) together in collaboration to develop and deploy a European radio navigation satellite system called Galileo.

Galileo development followed an iterative approach illustrated in Fig. 9.1. It was initiated in late 2003, carried out by the European Space Agency (ESA), and co-funded by ESA and the European Union.

ESA launched two GIOVE (Galileo In-Orbit Validation Element) satellites in 2005 and 2008, with a representative ground segment. These satellites secured the frequencies provisionally set aside for Galileo by the International Telecommunications Union. The satellites served also as a testbed for key technologies such as onboard atomic clocks and navigation signal generation. The GIOVE satellites are no longer active and have been moved to higher altitudes, away from the nominal Galileo orbit.

The following in-orbit validation phase aimed to perform initial validation of the system using a reduced constellation of four Galileo in-orbit validation (IOV) satellites – the minimum number for independent position and timing solutions at test locations – in combination with Galileo’s terrestrial network of ground stations. This phase used the first family of Galileo satellites (GSAT010x), launched through dual launches on 21 October 2011 and 12 October 2012. These four satellites served for IOV of the Galileo system, but are also part of the operational Galileo constellation. On 12 March 2013 this ground and space infrastructure came together to perform the very first determination of a ground location through Galileo signals alone. This initial position fix of longitude, latitude and altitude took place at the Navigation Laboratory at ESA’s technical heart European Space Research and Technology Centre (ESTEC), in Noordwijk, the Netherlands. From this point onward, Galileo navigation messages have

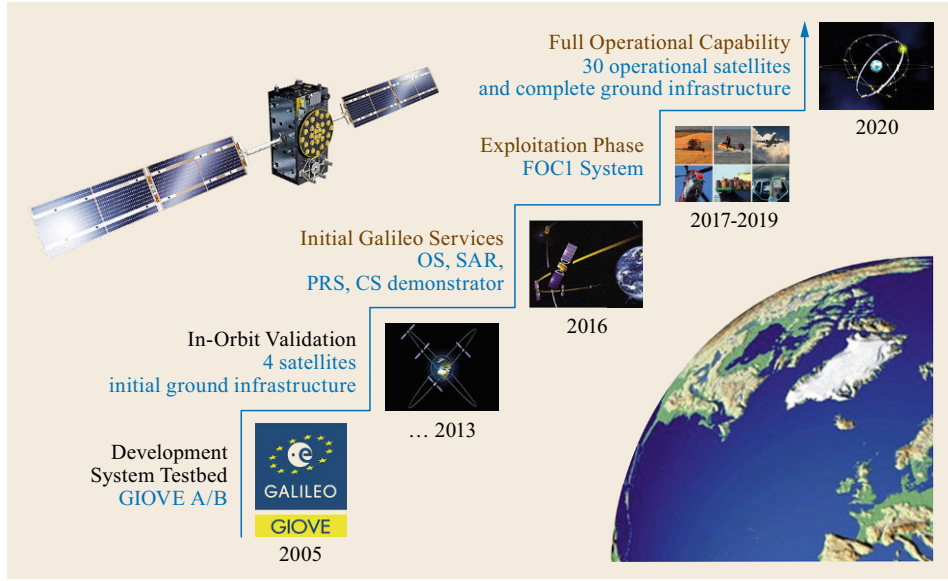


Fig. 9.1 Galileo incremental deployment (courtesy of M. Pedoussaut, S. Corvaja, Th. Burger, ESA, and pixabay.com)

been broadcast. The IOV campaign was conducted successfully throughout 2013, and the results were used as a reference to predict the expected performance of the completed Galileo constellation.

The deployment of the Galileo system until its full operational capability (FOC) is conducted under a public procurement scheme, entirely financed by the European Union. In 2007 the European Parliament and the European Commission decided to implement the system, and allocated budget for Galileo and for the European geostationary navigation overlay service (EGNOS). In 2008 the first part of the procurement of the Galileo full operational capability was launched, aiming to address full system deployment, long-term

operations and replenishment. This phase will consist of the launch of all remaining satellites (up to 30) and deployment of the full operational ground segment, including all required redundancies in order to comply with the full mission requirements in terms of performance and service area. Early Galileo services are set to begin during the year 2016.

Following this, the Exploitation Phase is planned to start during the deployment of the full system, scheduled for 2017, and will consist of routine operations as well as ground segment maintenance and replenishment of the satellite constellation. This phase is planned to last over the design lifetime of the system, nominally 20 years.

9.1 Constellation

The Galileo constellation is the result of detailed studies and optimization [9.1, 2]. Table 9.1 summarizes the finally selected basic Galileo reference constellation parameters [9.3, 4].

The nominal satellite positions in space for a given time are defined by the reference Keplerian elements expressed in the Celestial Intermediate Reference System (CIRS) [9.5].

$$\begin{aligned}
 i_{\text{ref}} &= 56^\circ, \\
 \Omega_{\text{ref}} &= \Omega_0 + 120^\circ \cdot (k_{\text{plane}} - 1) + \dot{\Omega} \cdot (T - T_0), \\
 u &= u_0 + 45^\circ \cdot (k_{\text{slot}} - 1) \\
 &\quad + 15^\circ \cdot (k_{\text{plane}} - 1) + D_{\text{nom}} \cdot (T - T_0). \quad (9.1)
 \end{aligned}$$

$$\begin{aligned}
 \Omega_0 &= 25^\circ, \\
 \dot{\Omega} &= -0.02764398^\circ/\text{d}, \\
 T_0 &= 21 \text{ March } 2010, \text{ } 00:00:00 \text{ UTC} \\
 u_0 &= 338.333^\circ, \\
 D_{\text{nom}} &= 613.72253566^\circ/\text{d}. \quad (9.2)
 \end{aligned}$$

In the equations above, the variable k_{plane} can take the values 1, 2 and 3 for, respectively, planes A, B and C [9.6]. The variable k_{slot} denotes the slots within the orbital plane and can assume values from 1 to 8. i_{ref} is the orbit inclination and Ω_{ref} the right ascension of ascending node (RAAN). The argument of latitude u is defined as along-track phase angle with respect to the equator.

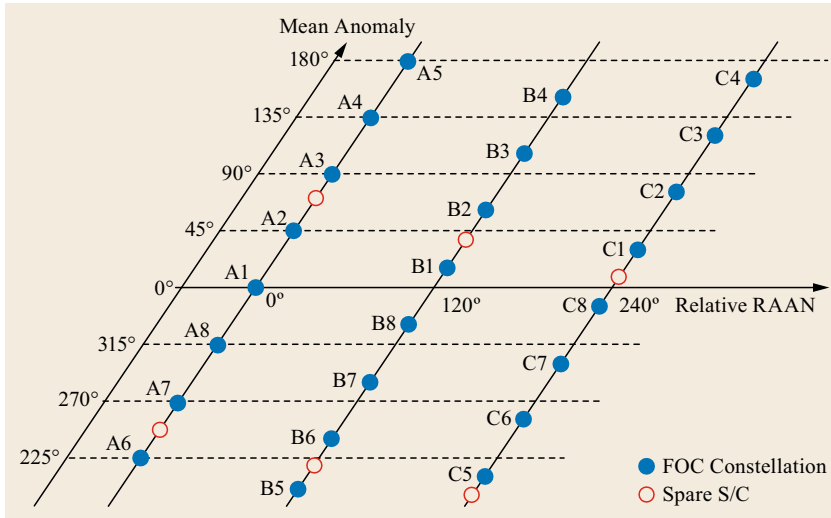


Fig. 9.2 Galileo FOC constellation slots

Table 9.1 Galileo reference constellation parameters

Parameter	Value
Reference constellation type	Walker 24/3/1 + 6 in-orbit spares
Semimajor axis	29 600.318 km
Inclination	56°
Period	14h 04m 42s
Ground track repeat cycle	10 sidereal days/17 orbits

The satellites are generally maintained within orbit slots of $\pm 2^\circ$ in inclination and argument of latitude, and $\pm 1^\circ$ in right ascension of ascending node (Ω) with respect to the reference. $\dot{\Omega}$ reflects both the oblateness of the Earth's gravitational field and the gravitational effect of the Moon and Sun. D_{nom} is computed taking into account the repeat ground-track pattern of 17 revolutions in ten sidereal days. Figure 9.2 shows a systematic sketch of the resulting constellation geometry and satellite locations.

The position of the spare satellites shown in Fig. 9.2 is indicative, as their actual position will be decided at the time of deployment.

The constellation geometry has been optimized to achieve consistently good geometric conditions on a global scale, leading to a good user position accuracy and availability. The inclination of the orbit planes provides for better coverage in the higher latitudes, for example when compared to GPS.

The above reference constellation of 24 satellites will yield between six to 11 Galileo satellites visible at any user location worldwide. Average visibility is more than eight Galileo satellites above 5° elevation. The reference constellation will be complemented with nominally six spare satellites. This constellation provides good local geometries with a typical vertical dilution of precision (VDOP, [9.7]) of 2.3 and horizontal dilution of precision (HDOP) around 1.3. An additional benefit of the constellation geometry is the limited number of planes, which allows for faster deployment and reduced constellation maintenance costs due to the capability to launch multiple satellites with a single launcher. For example Ariane 5 is capable of launching up to four Galileo satellites, and Soyuz is used to launch sets of two Galileo satellites.

Satellite disposal after the end of their operational life is to be considered and planned appropriately, for reasons of space debris control but also because debris avoidance maneuvers will impact the availability of operational satellites for service provision. Galileo satellites are removed from the nominal Galileo orbits after they have reached the end of their operational life. The same applies for remaining launcher stages after injection of new satellites into their orbits. The strategy followed in Galileo is to move those satellites and launcher stages to a *graveyard* orbit that is at least 300 km higher than the operational Galileo orbits.

9.2 Signals and Services

Each Galileo satellite provides coherent navigation signals on three different frequencies. Each signal contains several components, comprising always at least one pair of pilot and data components. Figure 9.3 summarizes the transmission plan.

The signals and components are assigned to three types of positioning services:

1. The Open Service (OS) comprising the data-pilot pairs E1-B/C, E5a-I/Q and E5b-I/Q, representing the publicly accessible positioning service
2. The Public Regulated Service (PRS) on E1-A and E6-A, a restricted access positioning service for government-authorized users
3. The Commercial Service (CS) through the data-pilot pair E6-B/C, a navigation signal on a third frequency, optionally encrypted, for the provision of future added value services.

As a fourth service, Galileo satellites support Cospas-Sarsat [9.8–10], an international satellite-based search and rescue system established by the US, Russia, Canada and France, capable of locating emergency radio beacons. This support is provided through a forward search-and-rescue repeater as part of the payload, and through an associated data return link embedded into the navigation message of the E1 OS data component.

E1 and E6 each provide one publicly accessible pair of pilot and data components, and E5 offers two pilot-data pairs in sidebands 15.345 MHz above (E5b) and below (E5a) the E5 carrier frequency. The sidebands E5a and E5b are foreseen for individual tracking and use, offering the equivalent of two coherent carrier frequencies within the E5 band. The overall E5 carrier including both sidebands is defined and generated coherently as an alternative BOC (AltBOC) signal [9.11]. This composite AltBOC signal can also be tracked as a single signal, offering a very large signal bandwidth of at least 51.15 MHz ($50 \cdot 1.023$ MHz), and thus providing excellent Gabor bandwidth (Chap. 4) and multipath rejection.

Table 9.2 gives the overview of available Galileo signals and associated carrier and subcarrier frequencies.

Galileo and GPS, as the system with the longest heritage and current most wide use, are considered compatible (sharing of resources without degrading the performance of the other radio navigation satellite system) and interoperable (allowing the user to successfully combine pseudorange measurements from more than one global navigation satellite system (GNSS) into position/velocity/time solutions):

1. Two carrier frequencies are shared (E5a/L5 and E1/L1), with equivalent modulations

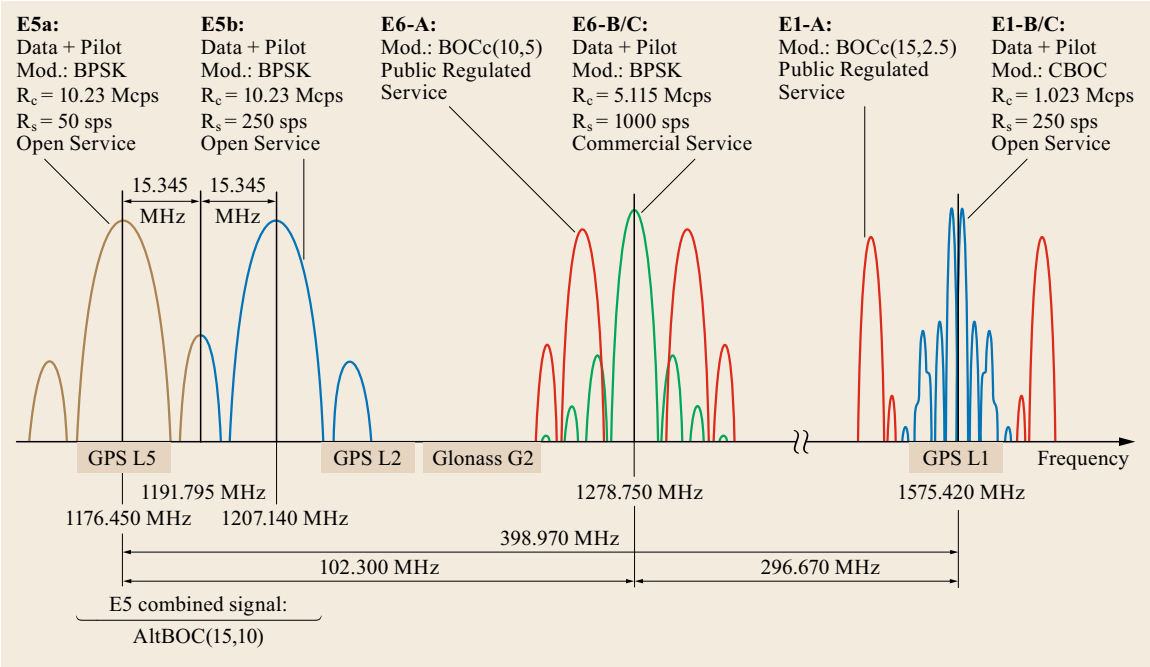


Fig. 9.3 Galileo frequency bands, signals and components

Table 9.2 Overview of Galileo signals

Galileo signal	Carrier frequency (MHz)	Subband	Subband frequency (MHz)	Carrier aligned with
E1	1575.420	n/a	n/a	GPS L1 C/A, L1C
E6	1278.750	n/a	n/a	
E5	1191.795	E5b	1207.140	
		E5a	1176.450	GPS L5

- Fundamental message concepts are comparable, such as ephemeris, almanac, clock correction, **GST-UTC** (Galileo System Time - Universal Time Coordinated), and bias group delay
- Terrestrial reference frames and reference time systems are aligned, as indicated in the following sections of this chapter.

These concepts and measures are intended to simplify and optimize user receiver hardware implementation (radio frequency front-end design and digital preprocessing), but also receiver software and algorithms.

9.2.1 Signal Components and Modulations

All Galileo signals and their signal components are derived from the same onboard master clock, and are thus coherent. Table 9.3 provides a summary of the modula-

tion schemes and component parameters. Four pairs of pilot and data components are provided for public use: E1-B/C, E6-B/C, E5a-I/Q and E5b-I/Q. All pilot/data pairs use a 50% power sharing.

The modulation-specific recommendations for receiver bandwidths are driven by the components to be tracked. Recommended receiver bandwidths are as listed in Table 9.4.

When choosing a receiver bandwidth, manufacturers are recommended to carefully consider the interference situation in each band. Some selected examples follow (the list is not complete). In E6 the presence of terrestrial pulsed interference, for example from radar systems, is to be expected, and the band is shared with amateur radio users that use it for audio and video transmitters and relays (HAM TV). Often also other unidentified low-power sources have been observed, especially in urban areas. Receivers will need to be resistant against high-power radio frequency (RF) pulses

Table 9.3 Overview of Galileo public signal components and modulations. Legend: R_c = primary code chip rate (in multiples of 1.023 MHz), R_{sc} = subcarrier frequency (in multiples of 1.023 MHz), R_d = symbol rate (symbols/s), R_{sec} = secondary code chip rate (chips/s)

Signal	Component	Modulation	R_c	R_{sc}	R_d, R_{sec}	Message	Service	Multi-plex	Min received Power
E1	E1-B Data	Composite binary offset carrier (CBOC) 1/11	1	1&6	250	I/NAV	OS	in	-160 dBW
1575.420 MHz	E1-C Pilot	CBOC 1/11	1	1&6	250	–	OS	phase	-160 dBW
E6	E6-B Data	Binary phase-shift keying (BPSK)	5	–	1000	C/NAV	CS	in	-158 dBW
1278.750 MHz	E6-C Pilot	BPSK	5	–	1000	–	CS	phase	-158 dBW
E5b	E5b-I Data	BPSK	10	–	250, 1000	[I/NAV]	OS	0°	-158 dBW
1207.140 MHz	E5b-Q Pilot	BPSK	10	–	1000	–	OS	90°	-158 dBW
E5a	E5a-I Data	BPSK	10	–	50, 1000	F/NAV	OS	0°	-158 dBW
1176.450 MHz	E5a-Q Pilot	BPSK	10	–	1000	–	OS	90°	-158 dBW

Table 9.4 Recommended receiver bandwidths for Galileo navigation signals

Component	Rx Bandwidth (double sided)			Note
	Min	Recommended	Max	
E1-B/C tracked as BOC(1,1)	2.0 MHz	2.0...24.6 MHz	≈31 MHz	In steps of 2.023 MHz
E1-B/C tracked as CBOC	14.3 MHz	14.3...30.7 MHz	≈31 MHz	Good multipath robustness already at 14.3 MHz
E6-B/C BPSK(5)	10.2 MHz	10.2...20.5 MHz	≈41 MHz	In steps of 10.23 MHz
E5a-I/Q BPSK(10)	20.5 MHz	20.5 MHz	≈41 MHz	In steps of 20.46 MHz
E5b-I/Q BPSK(10)	20.5 MHz	20.5 MHz	≈41 MHz	In steps of 20.46 MHz
E5 as AltBOC	51.2 MHz	51.2 MHz	≈72 MHz	Center frequency 1191.795 MHz

in-band, and against continuous transmitters close to band. The E5 band is mainly shared with air traffic control and positioning systems like DME (Distance Measuring Equipment, primary user), where the DME ground stations transmit within the E5 navigation band. DME transmissions are pulse pairs, each pair a few tens of μs long, transmitted at average rates up to several kHz and pulse powers in the range of 1 kW.

Selected Galileo Modulation Details

Galileo uses CBOC (Composite Binary Offset Carrier) modulation in E1 and AltBOC (alternative binary offset carrier) modulation in E5. These modulations are specific for Galileo and are shortly explained below. The Galileo E6 public signal uses conventional BPSK (binary phase shift keying) modulation as described in Chap. 4, and is thus not elaborated here. A full description of the public Galileo modulations is published in the Galileo public Open Service Signal in Space Interface Control Document (OS SIS ICD) [9.11].

CBOC as used for the Galileo E1 public signal is a composition of a 1.023 Mcps spreading sequence combined with a two-component spreading symbol. The spreading symbol comprises the sum of a BOC(1,1) subcarrier at 10/11 power and a BOC(6,1)

subcarrier at 1/11 power. The spreading symbol of the data channel combines the two subcarriers in-phase and the spreading symbol of the pilot channel combines them in antiphase, as illustrated in Figs. 9.4 and 9.5. As a result of the additive combination of the two binary offset carrier (BOC) subcarriers with nonequal amplitude, the time domain CBOC spreading symbols are four-level pulses. The spreading pulses of the CBOC data and pilot components differ with respect to the phase of their BOC(6,1) subcarriers.

When tracking using a conventional dual-level BOC(1,1) despreading is possible with minor losses ($\approx 0.4\text{ dB}$) that are also a function of the receiver bandwidth. Direct CBOC tracking requires a four-level correlator with amplitude stages of $\{\pm 1.25, \pm 0.65\}$. An approximation of the replica levels using two bit representations of the replica is possible, but not optimum. Alternative techniques, for example combining separate binary correlators for the BOC(6,1) and BOC(1,1) parts are increasingly becoming subject to publications and patents [9.12–14], demonstrating the feasibility of efficient tracking of CBOC modulated signals.

The phase of the BOC(6,1) components in the data and pilot spreading symbols is inverted. Thus the com-

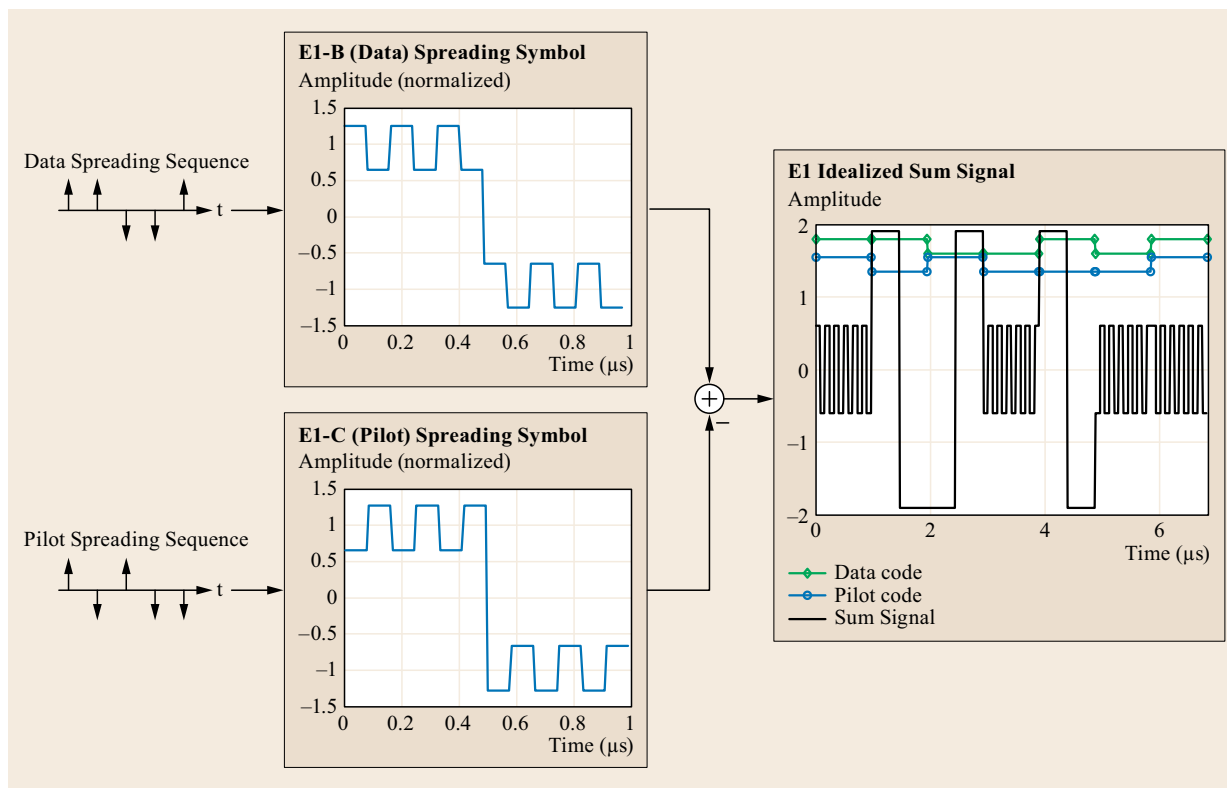


Fig. 9.4 Galileo CBOC principle

bined signal of pilot and data channel, as the in-phase addition of the pilot and data baseband signals, has only four levels in total. The combined signal has the interesting property that always only either BOC(1,1) or BOC(6,1) is transmitted, in time multiplex. The sub-carrier phase is set according to the combination of the spreading chips from pilot and data channels. This opens a range of possible highly efficient correlation mechanisms for tracking the combination signal, using time multiplex techniques.

AltBOC modulation was proposed in 2002/2003 [9.15, 16] as wideband complex sideband modulation. AltBOC can be understood in baseband representation as the sum signal of coherently generated and individually quadrature modulated complex upper (E5b) and lower (E5a) subcarriers, then adding an intermodulation function (IM) to achieve a constant envelope on the transmit side [9.17]. The OS SIS ICD [9.11] baseband representation

$$s_{E5}(t) = \sqrt{1/8} \cdot \begin{bmatrix} \begin{bmatrix} e_{E5a-I}(t) + j e_{E5a-Q}(t) \cdot \\ \left[sc_S(t) - j sc_S \left(t - \frac{T_s}{4} \right) \right] \\ + [e_{E5b-I}(t) + j e_{E5b-Q}(t)] \cdot \\ \left[sc_S(t) + j sc_S \left(t - \frac{T_s}{4} \right) \right] \end{bmatrix} & \text{Signal Part} \\ \begin{bmatrix} \bar{e}_{E5a-I}(t) + j \bar{e}_{E5a-Q}(t) \cdot \\ \left[sc_P(t) - j sc_P \left(t - \frac{T_s}{4} \right) \right] \\ + [\bar{e}_{E5b-I}(t) + j \bar{e}_{E5b-Q}(t)] \cdot \\ \left[sc_P(t) + j sc_P \left(t - \frac{T_s}{4} \right) \right] \end{bmatrix} & \text{Inter-modulation Part} \end{bmatrix} \quad (9.3)$$

contains in its first two lines the four independent bipolar $\{-1, +1\}$ spreading sequences $e_{E5\{a,b\}-\{I,Q\}}(t)$ (spreading code, secondary code and data modulation) in sideband modulation with their subcarrier $sc_S(t)$, and in the last two lines the IM consisting of the bipolar sequences $\bar{e}_{E5\{a,b\}-\{I,Q\}}(t)$ with their IM subcarrier $sc_P(t)$. All IM sequences are triple-product terms of the nominal spreading sequences $e_{E5\{a,b\}-\{I,Q\}}(t)$, for example

$$\bar{e}_{E5a-I}(t) = e_{E5a-Q}(t) e_{E5b-I}(t) e_{E5b-Q}(t).$$

The subcarriers before band limitation are discrete multilevel signals with period $T_s = (15.345 \text{ MHz})^{-1}$ and are defined as shown in Fig. 9.6.

The description provided in [9.11] is ideal wideband and yields a final signal constellation diagram (signal

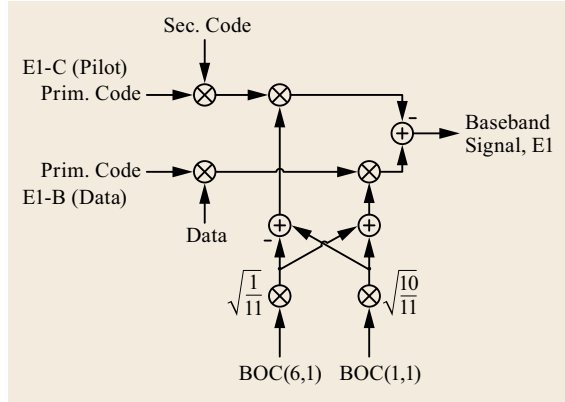


Fig. 9.5 CBOC generation block diagram

part combined with intermodulation function) that represents an eight PSK-type modulation (Fig. 9.7a). The main energy content of the IM is located around and beyond $\pm 46 \text{ MHz}$ offset from the E5 carrier (Fig. 9.7b), and lies outside the recommended AltBOC receive bandwidth (51.2 MHz). Receivers will see only a small fraction of the theoretical IM power and may thus safely decide to neglect the IM for the purpose of AltBOC tracking.

Various alternative AltBOC tracking concepts have already been published, and all require an AltBOC replica generation. One fundamental concept to generate the replica using a lookup table approach appears suitable for receiver implementation and is provided in [9.11], together with the direct mathematical description. This concept represents a baseline; it is to be expected that actual receiver implementations will use optimized forms, for example of combined replica generation and correlation computation.

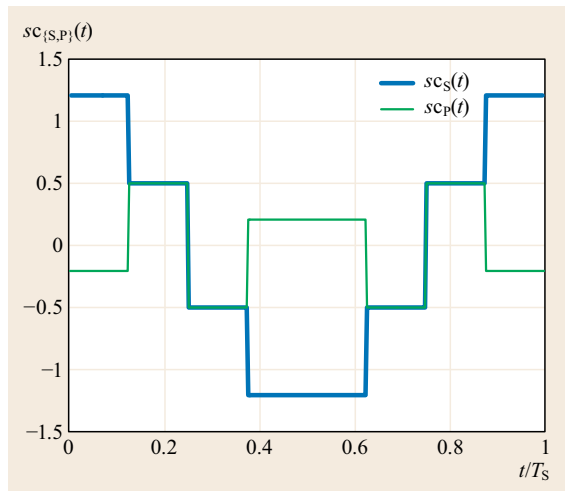


Fig. 9.6 AltBOC subcarrier functions

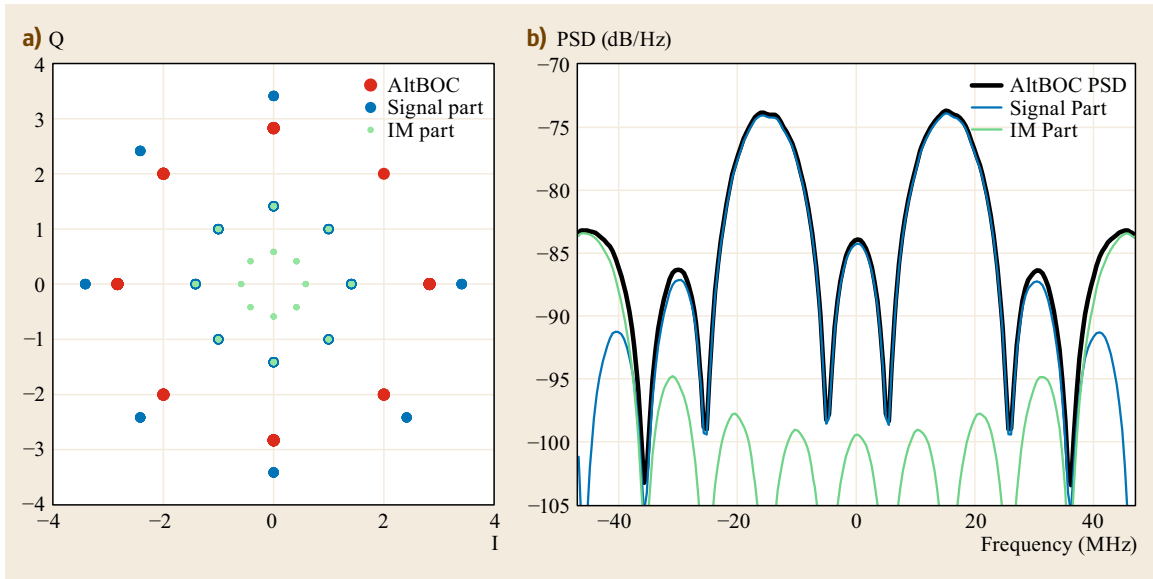


Fig. 9.7a,b AltBOC wideband signal vector diagram (a) and example power spectral density (b)

It is to be recalled that the Galileo navigation messages do not formally provide a direct message for use with AltBOC. The ephemeris information is unproblematic; any of the ephemeris sets provided in the Galileo navigation messages may be used. The clock correction is more critical, since the clock corrections as provided in the two public Galileo navigation messages are individual for specific frequency pairs (Sect. 9.2.2). If needed, either of these clock corrections and broadcast group delays (BGDs), or an average of them, may be used as good approximations.

In the case where only sidebands are to be tracked, then each of the four signal components $e_{E5\{a,b\}-\{I,Q\}}(t)$ can be acquired and tracked individually, as BPSK(10)-type navigation signal components. The two components on each sideband are configured as pairs of pilot and data. The two sidebands E5a and E5b of an AltBOC signal are fully coherent, thus any crosstalk between those sidebands is stationary. Tracking accuracy can suffer nonnegligible side effects from this

crosstalk. Individual components $e_{E5\{a,b\}-\{I,Q\}}(t)$ should thus be tracked using a receive bandwidth centered on the desired sideband E5a or E5b, and narrow enough to suppress the other sideband. For this reason Table 9.4 recommends component tracking bandwidths of 20.46 MHz.

Galileo Spreading Codes and Sequences

Each unencrypted signal component from each satellite is using individual, unique periodic spreading sequences (Table 9.5). The length (period) of the spreading sequences of data components is chosen such as to span full symbols of the data channel. If this requires more than 10230 chips, a two-tiered construction of a primary spreading code overlaid with a slower secondary code is used. The spreading sequences of pilot components generally use the two-tiered construction, with the length of the primary code equaling the primary code of the corresponding data channel, and the length of the secondary code chosen such as to pro-

Table 9.5 Overview of Galileo spreading codes (LFSR = linear feedback shift register)

Signal component		Primary code				Secondary code		
		Type	Chips	Period (ms)	#	Chips	Period (ms)	#
E1-B	Data, CBOC(1,6,1/11), 250 sps	Memory	4092	4	50	—	—	—
E1-C	Pilot, CBOC(1,6,1/11)	Memory	4092	4	50	25	100	1
E6-B	Data, BPSK(5), 1000 sps	Memory	5115	1	50	—	—	—
E6-C	Pilot BPSK(5)	Memory	5115	1	50	100	100	50
E5b-I	Data, BPSK(10), 250 sps	LFSR	10230	1	50	4	4	1
-Q	Pilot, BPSK(10)	LFSR	10230	1	50	100	100	50
E5a-I	Data, BPSK(10), 50 sps	LFSR	10230	1	50	20	20	1
-Q	Pilot, BPSK(10)	LFSR	10230	1	50	100	100	50

vide a total of 100 ms nonrepetitive length of the pilot spreading sequence.

The two-tiered spreading sequence generation works comparably to a pseudodata modulation, where the secondary code represents the (a priori) symbol modulation. Accordingly, the secondary code clocks with one chip per period of the primary code, and is modulo-2 combined with the primary code. Figure 9.8 illustrates this principle.

The design objective was to limit the length of the primary code to less than or equal to 10230 chips, to avoid excessive code search space during acquisition, but also to provide a nonrepetitive sequence length of either one symbol for data components or 100 ms for pilot components. Primary spreading sequences have been carefully selected and optimized for good orthogonality across each family, and are responsible for ensuring sufficient isolation between signal sources. Secondary codes have been mainly tuned for low autocorrelation sidelobes, and as a consequence, a flat power spectrum in the frequency domain.

The two-tiered construction cannot reach the correlation quality of an optimized single-stage spreading sequence with the same length as the combination of primary and secondary code. Instead, for coherent integration covering multiple primary code lengths, or over the full period of the tiered code, the correlation result will repeat the primary code autocorrelation modulated with the shape of the secondary code (partial) autocorrelation. This means there will be repeating correlation peaks every period of the primary code, not with the full amplitude of the main correlation peak, but still

with significant levels. Receivers using coherent integration times longer than the primary code period will need to consider this behavior through appropriate hypothesis tests during acquisition to find the correct main peak and secondary code phase. But once the phase of the secondary code is identified, this resolves the code phase unambiguously up to the length of the tiered code. The two-tiered code concept with its pseudodata modulation in the form of the secondary code is also intended to reduce the sensitivity to narrowband interference, compared to repetitive primary codes without secondary code, while maintaining a reasonably length-limited primary code.

The secondary code on pilot components allows the resolution of the code phase relative to Galileo System Time (GST) with an ambiguity interval of 100 ms. This approximately equals the maximum propagation delay between any visible Galileo satellite of the nominal constellation and terrestrial users, and is more than four times the difference between the propagation delay to the closest and the farthest user on Earth. It is therefore considered possible to derive time-free position solutions for users on the surface of Earth, using only code phase measurements including the secondary code of any pilot signal, and provided the receiver already has ephemeris and clock correction information available.

The Galileo primary and secondary spreading codes for public use are provided in the OS SIS ICD [9.11]. Note that the memory codes are provided only in the downloadable electronic (PDF) version of that document. The occasionally available paper printed version

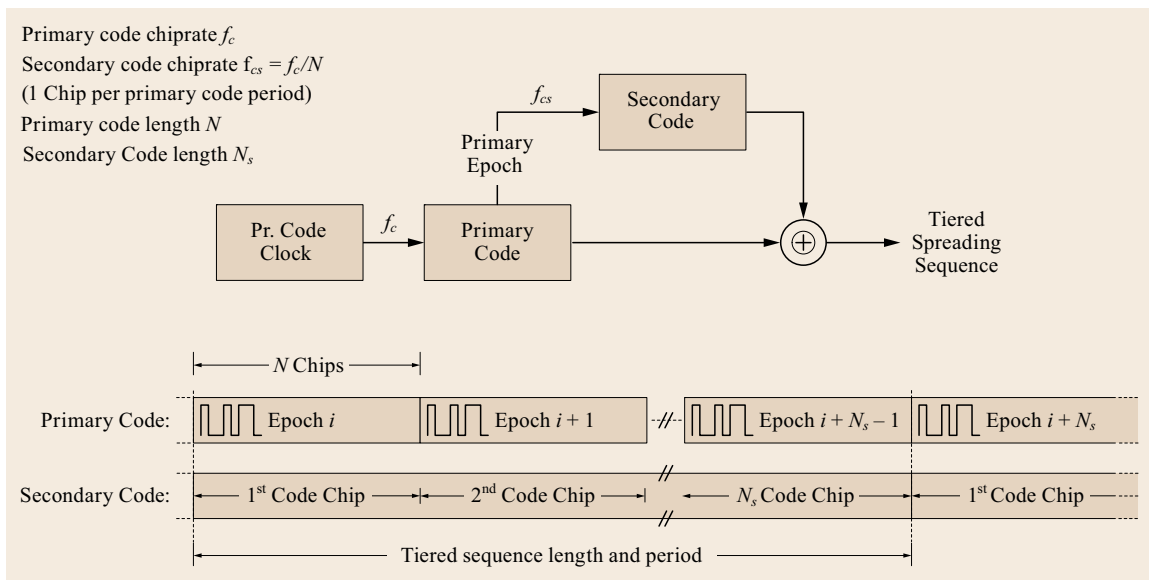


Fig. 9.8 Principle of the tiered code construction

may not contain hexadecimal representations of the memory codes.

9.2.2 Navigation Message and Services

Three different types of public navigation messages are provided through the Galileo navigation signals: the high data rate and short page length I/NAV (historically derived from Integrity NAVigation message), the low data rate F/NAV (Free NAVigation message) and the fast C/NAV Commercial channel NAVigation message. The message types are assigned to signal components as described in Table 9.6. The OS SIS ICD [9.11] and its annexes and associated support documents serve as reference documentation for these message types. OS SIS ICD [9.11] will be gradually extended and amended with new content, following service deployment and the progress of system validation. This section will thus reference to [9.11] and otherwise focus on receiver relevant differences and specifics.

The content of the navigation messages can be roughly differentiated into position/velocity/time (PVT) relevant content, which is mostly repetitive, and nonrepetitive low latency message elements.

Both I/NAV and F/NAV provide direct support to PVT determination, through provision of GST in the form of week number (WN) and time of week (TOW), and of ephemeris and clock correction for the transmitting satellite, but also through ionosphere model parameters, bias group delay information needed for single-frequency users, data validity and signal health flags, almanac and other supplementary information. The fundamentals of ephemeris, clock correction, GST-UTC, almanacs and usage algorithms are consistent with GPS legacy definitions, with format adjustments to Galileo. The ionosphere correction message uses a Galileo-adapted version of the more recent NeQuick model. The detailed user algorithm reference model will be published as annex [9.18] to [9.11].

For low-latency content, I/NAV on E1 includes the return link channel supporting the Cospas-Sarsat search-and-rescue (SAR) system [9.8–10]. This return link is a near-real-time channel for short messages to SAR beacons equipped with a Galileo navigation receiver. Further low latency channels are embedded into the I/NAV message but are not yet formally published,

being considered a functional reserve of the Galileo navigation message for future development. In case such low-latency channels are influencing and altering the flow of the message data stream, these changes need to be known to receivers and to be considered already now. One example is I/NAV on E1 and on E5b, which includes capabilities to replace nominal transmissions on a per second basis with one-time low-latency short message pages [9.11]. Despite these messages not being brought into use yet, receivers will need to be robust against such insertions as they may appear in future.

The C/NAV data stream on E6 is also implemented as a near-real-time message stream with short latency. At the time of writing C/NAV applications are under development [9.19] and no content is published yet.

All low-latency data channels are served only from satellites with active uplink from the Galileo ground segment. Their data content can differ between different satellites.

The content of I/NAV and F/NAV messages is compatible with almanac, ephemeris information, GST-UTC and GST-GPS time conversions. Clock correction parameters of I/NAV and F/NAV messages are specific per message type, and are expected to be very similar but are not guaranteed to be identical. This difference is the result of Galileo being a native multifrequency system, where both I/NAV and F/NAV messages are being optimized for specific pairs of frequencies. The I/NAV message, especially its clock correction, is calculated for dual-frequency reception of E1 and E5b, and F/NAV is optimized for the dual-frequency reception of E1 and E5a. As a result the ephemeris and clock corrections provided in F/NAV and I/NAV messages are directly applicable for dual-frequency receivers of the above frequency pairs. Any single-frequency receiver needs to use the bias group delay correction provided within the assigned message type to adjust the clock correction for the single frequency to be measured. Figure 9.9 illustrates this rule.

None of the messages published so far supports PVT using E6 measurements or triple-carrier measurements. Such content could be envisaged to be provided in the C/NAV message as future services or through external sources and communication channels.

Table 9.6 Message content coverage

Message type	Component	Content Positioning	Search and Rescue	Supplementary
F/NAV	E5a-I	✓		
I/NAV	E1-B	✓	✓	Individual low latency content
	E5b-I	✓		Individual low latency content
C/NAV	E6-B			C/NAV low latency content

Message Timeline and Structure

All Galileo message streams are structured in *pages* as the smallest interpretable data blocks. For F/NAV and C/NAV each page consists of a predefined set of synchronization symbols followed by the rate-1/2 convolutional encoded and cyclic redundancy check (CRC)-protected block of information. One F/NAV page lasts for 10 sec and provides 238 bits of effective information, exclusive of sync and tail symbols. I/NAV transmits data pages in two consecutive blocks, namely the *odd* and *even* words. Each word starts with the I/NAV synchronization symbols followed by a block-encoded data field, and lasts one second. A full I/NAV page (*odd* and *even* words combined) takes two seconds for transmission and provides a usable capacity of 245 bits, exclusive of sync and tail symbols.

The sequence of pages as transmitted from each satellite is organized such that the information required for PVT is provided within a well-defined maximum interval of time. Parts of information with less or no direct relevance for PVT and with longer validity, for example almanacs, are distributed over longer intervals. Figure 9.10 illustrates the concept for F/NAV.

It needs to be noted that [9.11] provides these structures for information with several reservations on possible changes and evolutions, intended to preserve some space for possible future improvement and extensions of the navigation message. Modulation up to

symbol level will not change, and existing pages will not disappear. Backward compatibility for legacy receivers will be maintained. But new features may be gradually introduced, using the available degrees of freedom and spare room of parameters like page type identifiers. User receiver designers are asked to consider these reservations appropriately. Some examples are:

1. The nominal sequence of pages as described in [9.11] is not to be relied upon, but may change in the future. This implies that the page sequence may not be the same for all active satellites within a Galileo constellation. The receivers need instead to identify received pages by their page type identifier.
2. The relative timing between I/NAV pages in E1 and in E5b may change, for example through the above changes of page sequences.
3. New page types may be introduced. This will not degrade the PVT quality achievable with legacy data content. However, receivers are expected to react to page types not known to them in a well-controlled manner. Similarly, other spare space in identifier value ranges may be explored and combined with new definitions of data content, for example almanacs for space vehicle identifiers (SVIDs) outside the currently defined range.

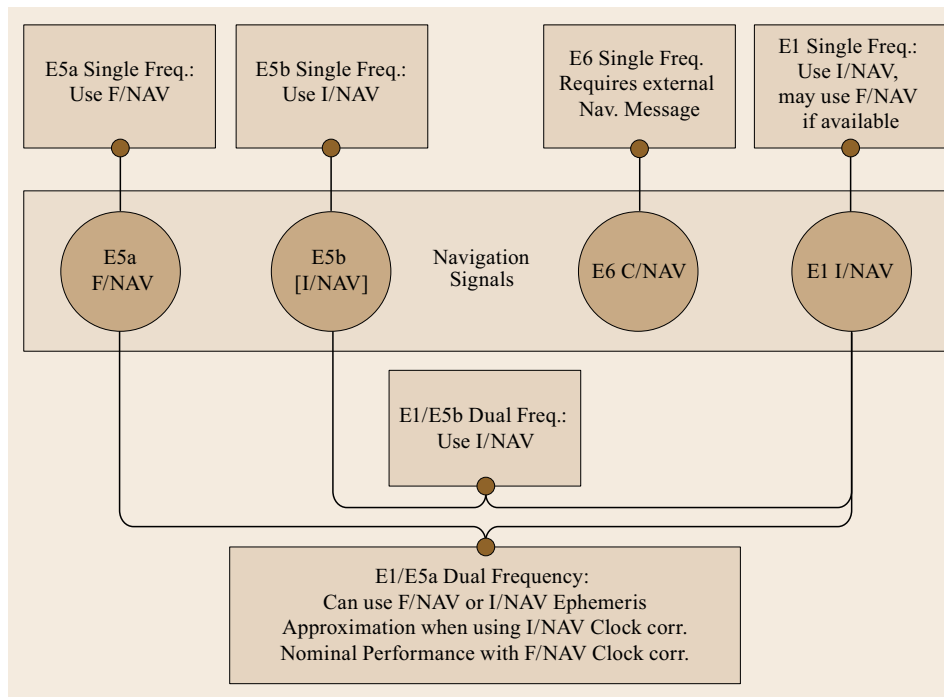
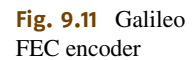
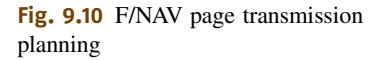


Fig. 9.9 Rules of use of Galileo navigation messages for PVT



- as independent data blocks without overlaps with earlier or later blocks.

Each block is subject to block interleaving using blocks with eight rows and a number of columns according to the page size in symbols (Fig. 9.12), supporting the textual representation in [9.11].

This ensures that burst errors of the channel are de-interleaved to at least eight symbols distance between single symbol errors at the decoder input, supporting the FEC decoder to correct such errors.

9.2.3 Ranging Performance

The Galileo ranging performance and consequently also positioning performance is driven by three groups of error contributions originating from the Galileo system, the environment and the user receiver.

The Galileo user equivalent range error (UERE) budget considers all key contributors as a function of satellite elevation:

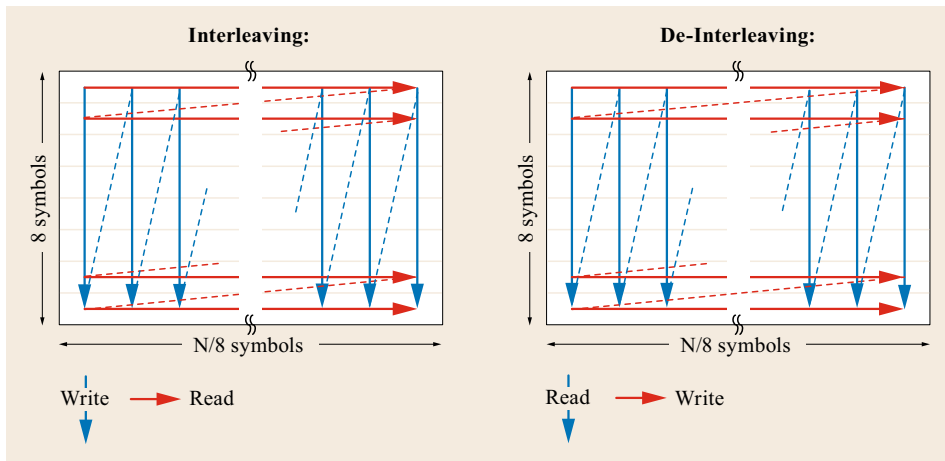


Fig. 9.12 Galileo message interleaving and de-interleaving

1. Ionospheric error: residual error due to the imperfection of the ionosphere model as provided in the navigation message, used to correct for ionospheric delays (only for single-frequency users)
2. Tropospheric error: residual error due to the imperfection of the model used to estimate the tropospheric delays. Models like Saastamoinen [9.20] combined with International Telecommunication Union (ITU)-R P.835-3, -4 will easily fulfill the assumptions reflected in Table 9.7
3. Interference, multipath, receiver thermal noise error: errors in the user receiver equipment due to *local* effects on code error, such as thermal noise, radio frequency interference and multipath
4. Orbit determination and time synchronization error: error caused by imperfections in the system-provided reference data (ephemeris and clock correction) for the computation of the satellite orbits and clocks at user level
5. Broadcast group delay (BGD) error: residual error due to the imperfection of the correction for transmitter delay differences between carriers (only relevant for single-frequency users).

Table 9.7 provides indicative root mean square (RMS) magnitudes of the error contributions expected to be achieved for the Galileo Open Service (OS) once the system has reached full operational capability. The values are valid for satellites at medium elevations around 45 degrees, at the maximum operational age of data.

The total UERE for a specific user and its environment combined with the local reception geometry, expressed for example through the dilution of precision (DOP), can be used to estimate the user position accuracy.

Table 9.7 Indicative root mean square (RMS) magnitudes of the Galileo user equivalent range error (UERE) contributions

UERE Contributor	Single Frequency user	Dual
Residual ionosphere error	< 500 cm	≈ 5 cm
Residual troposphere error	< 50 cm	< 50 cm
Thermal noise, interference, random multipath and multipath bias error	< 70 cm	< 100 cm
Orbit determination and time synchronization error	≈ 65 cm	≈ 65 cm
Satellite broadcast group delay	≈ 35 cm	0
Total (RMS)	< 513 cm	< 130 cm

Galileo comprises a worldwide network of initially 16 sensor stations to collect the ranging measurements required to generate the navigation message. The number of stations and their distribution accounts for eventual temporary sensor station and network outages. This ensures the required level of ground network robustness and performance as needed for the provision of nominal ranging, position and timing accuracy.

The Galileo system has been designed with position accuracy targets for E1/E5 dual-frequency Open Service users set to 4 m (95%) horizontal and 8 m (95%) vertical. The ranging accuracy considered necessary to reach these accuracy targets is 130 cm (95%). These targets are used to benchmark the performance predictions and to determine the expected availability of the service accuracy. Figure 9.13 provides global-scale simulated *expected performance* for an Open Service dual-frequency Galileo-only users in the vertical and horizontal position domain with 99.5% availability according to the above benchmark thresholds.

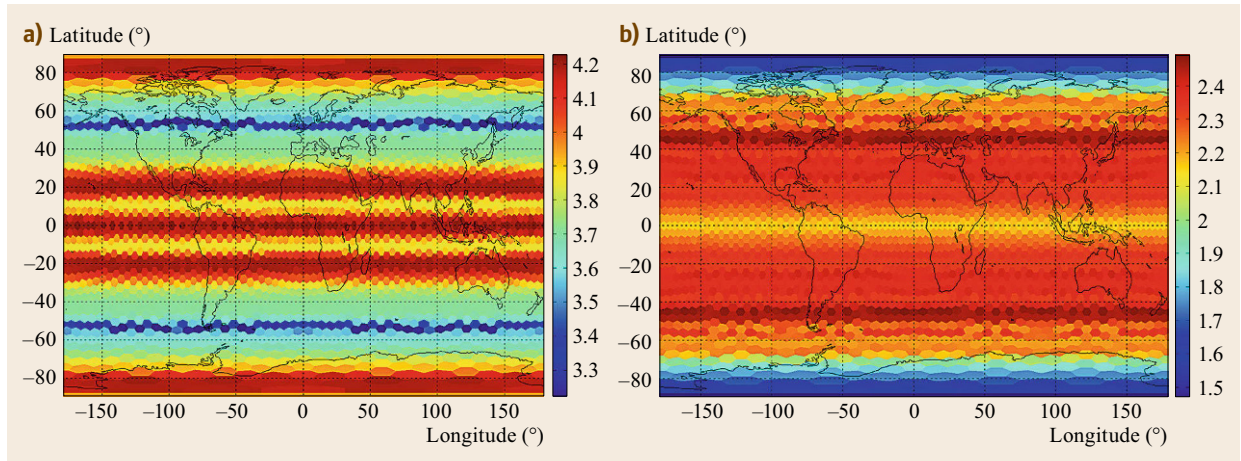


Fig. 9.13a,b Simulation of expected performance (color-coded positioning error in meters) for an Open Service dual-frequency Galileo-only user in the vertical (a) and horizontal (b) position domain with 99.5% availability

Orbit and Clock Errors

As for most radio navigation satellite systems, the basic information generated by the Galileo system for provision to users are orbit and clock correction for each satellite. The mission ground segment *estimates*, *predicts* and *parametrizes* this information into the navigation message. The message is then *uplinked* to the satellites and *disseminated* to the user through the navigation signals.

Clock and orbits are *estimated* by the orbit determination and time synchronization (ODTS) process, which operates as a least-squares estimator on data batches, operationally running every 10 min. Observation data used for this estimation process are always dual-frequency pairs of measurements:

- E1-E5a observables for F/NAV products
- E1-E5b observables for I/NAV products
- E1-E6 observables for PRS products.

Once estimated, clocks and orbits are *predicted* for the time interval for which the navigation message is to be generated. The reference time of these predictions is located at the beginning of each prediction interval. The result is subsequently *parameterized* and formatted into navigation messages, and finally disseminated to the user. Users will observe a difference (*age of data*) between reference time and time of use of the message. The signal-in-space error (SISE) at the user age of data, that is the imperfection of clock and orbit parameters when applied by the user, determines the ranging performance.

Stability and predictability of the satellite payload implementation, and especially the onboard clock system, is a major contribution to SISE. For this reason the quality of clock and orbit estimations is validated

systematically. A specifically interesting time for such measurements are the in-orbit tests (IOTs) following each satellite launch. During these periods all clocks on board can be operated, even if only for a limited time, and can be observed.

A selection of typical results for RMS clock prediction errors (clock predictability) is shown in Fig. 9.14, as a function of the prediction interval. The prediction model used is a second-order polynomial, equivalent to the clock correction within the navigation message, but in high numerical resolution and therefore without parameter quantization effects. Figure 9.14 is intended to show clock quality as much as possible isolated from other contributions. Therefore the clock estima-

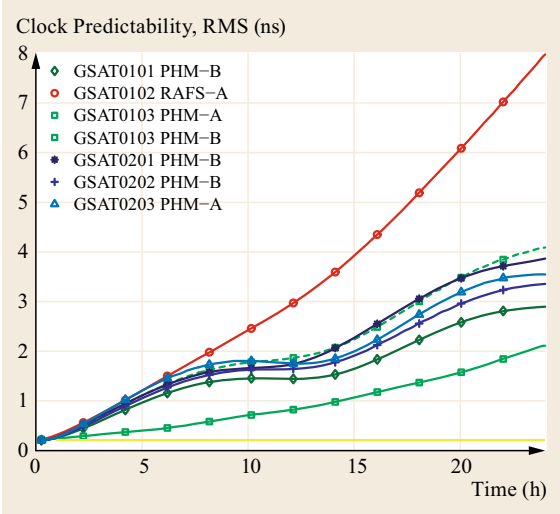


Fig. 9.14 RMS clock predictability for typical Galileo passive hydrogen maser (PHM) and RAFS

tion used originates from an independent verification system based not only on Galileo sensor stations but also including other project internal sensor stations, otherwise used for verification purposes, and sensor stations of the Multi-GNSS Experiment (MGEX), which is maintained by the International GNSS Service (IGS). A highly stable active maser independent of the Galileo system is used as ground time reference. The vertical bias for very short prediction intervals indicates the measurement noise of the input data as used for the analysis and the resulting model fitting error. Only one rubidium atomic frequency standard (RAFS) is shown, but its behavior is quite representative for the typical predictability of the GSAT010x and GSAT02xx RAFS. The difference between RAFS and passive hydrogen maser (PHM) predictability is clearly visible, especially for longer prediction intervals. The PHM quality and predictability is such that orbit determination imperfections can still be recognized, despite the offline processing using a larger database than available to the system itself.

Each satellite operates two clocks in parallel, nominally one PHM as a master clock and one RAFS as a hot redundant backup. In normal operation the ground segment commands switches between clocks, for example due to maintenance or, eventually, due to failures. Expectation is that at the time of switching to a backup clock, this clock has been in operation already for a prolonged time as a hot redundant clock. Thus its performance will have stabilized and settled, and the switch can be performed with minimum impact on service provision. Planned switches have meanwhile been exercised seamlessly, that is without interrupting service provision from the affected satellite.

Following established practice the ephemeris and clock correction message is generated with respect to a common reference point (apparent center of phase) geometrically close to the frequency-dependent individual centers of phase of the navigation transmit antenna. The user computes the position of this common reference point as a function of GST [9.11]. The vector from satellite center of mass to antenna reference point will be considered for possible publication, following validation.

Galileo establishes its own Galileo Terrestrial Reference Frame (GTRF), to which the above orbit and clock errors are referring. The GTRF is aligned to the International terrestrial reference frame (ITRF) with respect to origin, scale, orientation and rate, such as to remain within 3 cm (2-sigma) of the ITRF.

Ionospheric Error and Broadcast Group Delay

The ionospheric model parameters include the broadcast coefficients a_{i0} , a_{i1} , and a_{i2} used to compute the

effective ionisation level A_z , and the *ionospheric disturbance flags* (also referred as *storm flags*), provided for five different regions.

The ionospheric algorithm for single-frequency users is based on an adaptation of the three-dimensional (3-D) empirical climatological electron density model NeQuick [9.21–23].

The performance of the Galileo NeQuickG model is regularly evaluated. An early but still valid result was measured during the IOV campaign (March to August 2013) [9.21, 24]. The achieved residual error as measured was already reaching expectations for the full operational constellation of Galileo, and was better than the GPS Klobuchar model especially at equatorial latitudes. The global absolute ionospheric error ($1 - \sigma$) for the reported period was 1.34 m RMS achieved with the NeQuickG model and 1.9 m RMS with Klobuchar model. The absolute difference at equatorial latitudes expands to well beyond 1 m.

An example level of correction and comparison of NeQuickG and Klobuchar model results is provided in Fig. 9.15 for 21 May 2015, close to the vernal equinox and the seasonal ionosphere maximum. The performance was measured with receivers in the marked locations (more than 100 stations). From white to green the achieved ionospheric error correction capability was at least 70% RMS or better. Red color markings indicate a correction performance lower than 70%. This result is computed following the description in [9.25], and is in line with these earlier observations.

Similar to other GNSSs, the Galileo clock corrections are generated for dual-frequency users, and single-frequency users will need to use the broadcast group delay $BGD(f_1, f_2)$ provided through the Galileo navigation message as an additional correction. $BGD(f_1, f_2)$ is defined as follows

$$BGD(f_1, f_2) = \frac{TR_1 - TR_2}{1 - \left(\frac{f_1}{f_2}\right)^2}, \quad (9.4)$$

where f_1 and f_2 are the carrier frequencies of the involved Galileo signals 1 and 2, while $TR_1 - TR_2$ is the delay difference of the signals as contributed by the satellite payload. This formulation allows for easy translation of the dual-frequency clock correction information from the navigation message when using only a single-frequency receiver [9.11]. BGD accuracy was characterized, for example during the IOV campaign, and was found to be as expected around 30 cm. It is noted that BGD does not distinguish between pilot and data components. The ground segment measures BGD on the pilot components of the associated dual-frequency combination. The data components are

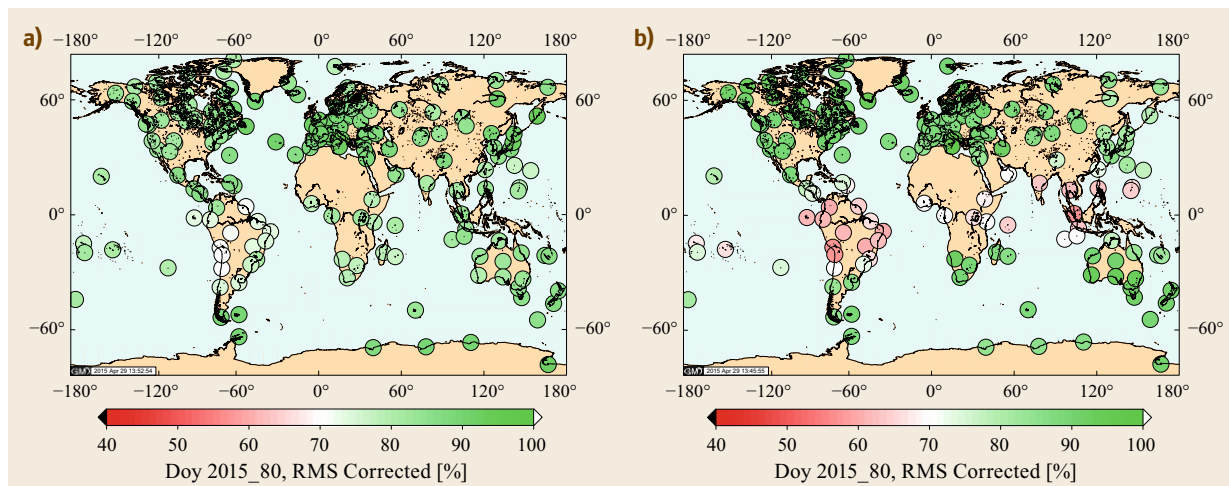


Fig. 9.15a,b Level of ionospheric correction and comparison of Galileo NeQuick G model (a) versus GPS Klobuchar model (b) for 21 May 2015

nearly identical to their pilot counterparts, in spectrum and modulation, and the method of signal generation on board ensures that tracking offsets between a data component and its pilot counterpart remain within a few cm, an order of magnitude smaller than BGD accuracy.

Message Uplink and Dissemination

The navigation message information is routinely generated by the orbitography and synchronization processing facility (OSPF) at intervals of 10 min. To decrease parameterization and quantization errors, the navigation message is generated in sets of eight batches, each batch marked through individual and unique issue of data (IOD) values. Subsets consisting of the first four batches are uploaded through mission uplink contacts and stored on board, using always the latest available sets. During the duration of mission uplink connections to each satellite, the message information broadcast from this satellite will thus also update approximately every 10 min, and the user is provided with the most recent and up-to-date navigation information. Between mission uplinks the satellites operate on the sets stored on board, broadcasting each message batch until it reaches a configured age of, for example, 180 min. Then the satellite selects the next message batch for dissemination, and will so step through the message batches in storage. In this example the broadcast of batches 2, 3, and 4 would start 3, 6, and 9 hours after the last mission uplink.

Navigation message batches can also be uplinked via telemetry, tracking and commanding (TT&C) stations. These uplinks can provide all batches 1–8 of a set, which allows longer operation on stored batches. In the above example, batches 5–8 would become valid 12,

15, 18, 21 hours after the last uplink contact. If there are no uplinks for intervals longer than the example 21 hours, the last message will be continuously transmitted regardless of its age.

In nominal operation the mission uplinks to the Galileo satellites will be scheduled tightly, for example such that the age of the onboard navigation message does not exceed 100 min, in support of using RAFS as master clocks. While using PHM master clocks the time between uplinks can then be extended.

These constraints need to be taken into account for sizing and location of the uplink station network. Accordingly, plans are for at least five uplink stations, where each station can operate up to four uplink antennas. This allows the supply of the complete Galileo constellation with timely navigation message data.

The broadcast of valid Galileo navigation messages started 12 March 2013, using the first four Galileo satellites, and then allowed for the first Galileo autonomous position fix [9.26].

Open Service Position and Ranging Error Accuracy

Position accuracy is monitored regularly. Receiver determined position solutions at user level, obtained in 2013 during periods with at least four Galileo satellites in visibility, are reported in [9.24], and since then the continuous deployment of the constellation and ground segment has led to the expected improvements. Position solutions collected in February 2016 are shown in Fig. 9.16, covering one 10 day ground track repeat cycle of the Galileo constellation. The position fixes were achieved through a dual-frequency E1b-E5a receiver

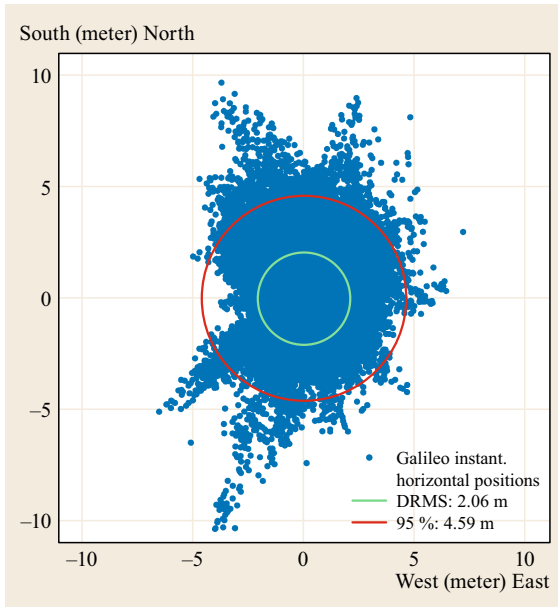


Fig. 9.16 Horizontal position accuracy at end-user level, Noordwijk, February 2016

where the geometry has been constrained for a geometric dilution of precision (PDOP) better or equal to 5. The measured horizontal accuracy shown in this example is well within expectations, exhibiting here less than 5 m 95% (green circle in Fig. 9.16).

These position results include the effect of dilution of precision, which is still not at nominal level due to the limited number of satellites available at the time of measurement.

The ranging accuracy is a figure of merit describing system performance per signal, before dilution of precision comes into effect. For example, SISE is derived by combining clock and orbit errors projected into the user direction. Its history during the project phases reflects the progress of system deployment and tuning. During IOV in 2013 only four satellites were available, and the ground segment was using only a subset of sensor stations. This initial validation achieved SISE results around 1.3 m 67%. In 2014 the SISE performance was approximately 1.0 m 67%, while in 2015, following the update of the ground segment, a SISE of 0.69 m 67% has been reached.

9.2.4 Timing Accuracy

The Galileo internal reference time is the Galileo System Time (GST), and is linked to Universal Time Coordinated (UTC). The navigation message provides GST-UTC information to allow estimation of UTC as an international time reference. To support interoper-

ability with GPS, Galileo also provides the measured GPS-to-Galileo time offset (GGTO).

System Time Generation

GST is a continuous coordinate time scale in a geocentric reference frame, steered towards UTC modulo 1 second. GST is not subjected to leap seconds.

GST is used as reference time throughout all Galileo system facilities, ground station and satellite clocks.

The broadcast navigation message is time-tagged with GST and provides GST as a 32-bit binary field composed of Galileo week number (WN) and time of week (ToW) [9.11].

The initial epoch of GST is defined to be 00:00 UTC on Sunday 22nd August 1999. This corresponds to the last rollover of the GPS week number. The offset between GST and UTC at the initial epoch is defined as 13 seconds. The GST-UTC offset is changing with the introduction of new leap seconds. This definition is a contribution to GPS-Galileo interoperability insofar as GST has zero integer seconds offset to GPS Time.

The Galileo Time Service Provider (GTSP) links GST to UTC, through selected European timing laboratories, and provides the required frequency steering correction such that GST can be steered to UTC within 50 ns 95% modulo 1 s. The GTSP also provides the GST-UTC offset and the leap second announcements.

GST is generated by the Galileo Ground Mission Segment based on the atomic clocks of the Galileo Precise Time Facility (PTF). Galileo uses two redundant PTFs, one in each of the two control centers in Fucino, Italy and in Oberpfaffenhofen, Germany. Each PTF is equipped with two active hydrogen masers in hot-redundant configuration, and with four high-performance Cesium clocks. The physical realization of GST is defined as the maser output, steered to UTC modulo 1 second according to the GTSP steering corrections.

The quality of GST steering towards UTC is demonstrated for example through a collaboration with several European timing laboratories, comprising Istituto Nazionale di Ricerca Metrologica (Italy), National Physical Laboratory (UK), Observatoire de Paris (France), Physikalisch-Technische Bundesanstalt (Germany), Real Instituto y Observatorio de la Armada (Spain), and Swedish National Testing and Research Institute (Sweden). Already the early verification in 2013 demonstrated very good alignment and stability [9.27]. The subsequent monitoring confirms this steering quality, during all deployment stages of the related subsystems. Figure 9.17 shows an example measurement covering the period from 1 January 2014 to 31 May 2014. In this period the offset UTC – GST be-

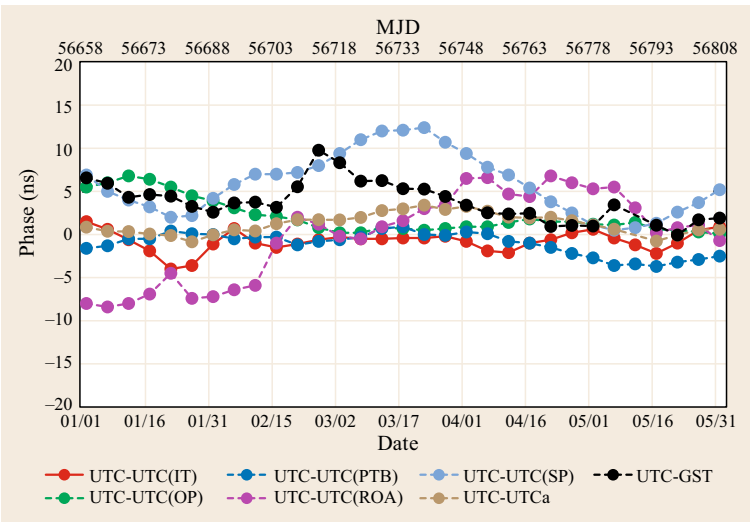


Fig. 9.17 GST versus UTC offset, relative to European timing laboratories; results from January to May 2014

tween the GST physical realization and UTC remains well within 10 ns, and the behavior of GST relative to UTC is quite similar to the evolution of the UTC offsets of the involved timing laboratories.

UTC Dissemination

In accordance with ITU Recommendation TF.460-6 [9.28], Galileo disseminates Universal Time Coordinated. For this purpose the Galileo navigation message contains GST-UTC conversion parameters including the total number of leap seconds (i.e., GST-UTC integer offset), announcement of introduction of new leap seconds with the associated date, fractional GST-UTC offset, and slope. Galileo users estimate GST from the broadcast navigation signals, then use the GST-UTC conversion parameters from the navigation message to estimate UTC for their applications. The GST-UTC conversion parameters are generated by the GTSP and updated daily.

The achievable Galileo UTC dissemination performance is being monitored by the Timing Validation Facility, a test and validation facility developed by ESA. Figure 9.18 shows the GST offset to UTC as disseminated by Galileo (UTC(SIS)) versus the rapid UTC solution, an official product of the Bureau International des Poids et Mesures (BIPM) closely approximating UTC. During the reported period May to September 2015 the UTC dissemination error was already well within expectations, but shows minor side effects and improvements in stability and offset due to ongoing calibration and equipment deployment.

GGTO Dissemination

GPS and Galileo system times are derived independently from each other. Both are kept aligned to UTC

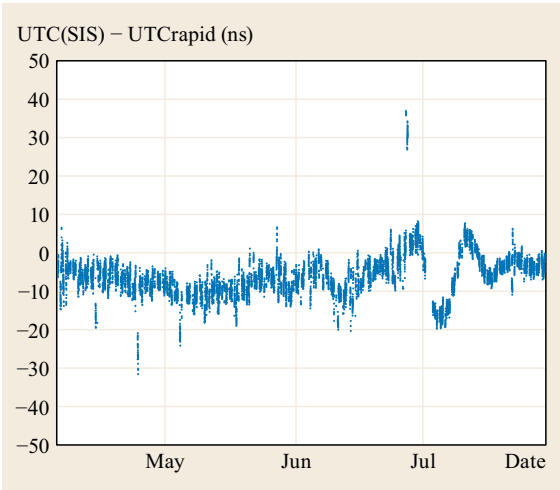


Fig. 9.18 Offset of UTC as disseminated by Galileo (UTC(SIS)) versus rapid UTC solution in 2015

within their respective GNSS. The Galileo PTF measures the offset between both system times and determines the GPS-to-Galileo time offset (GGTO), defined as the difference between the Galileo and GPS timescales, $GGTO = t_{Galileo} - t_{GPS}$ [9.29–31]. GGTO is then distributed through Galileo’s navigation message [9.11] to support combined use of Galileo and GPS systems. The provision of GGTO is expected to benefit user receivers in situations with obstructed field of view and limited numbers of visible satellites from a mixed constellation, can support integrity monitoring, and can support receiver internal calibration.

GPS has been chosen as the reference due to being the best established GNSS at the time. Currently Galileo does not provide offsets to other GNSS than

GPS. If all GNSS would provide their time offset relative to the same common reference, then a receiver could determine all required offsets for using any combination of measurements from mixed constellations for PVT computation. There would be no functional need to provide multiple offsets to different GNSSs within each navigation message. However, if this common reference would fail, all related and derived offsets may degrade or become unavailable. If the approach of a common reference were chosen, at least two independent references are recommended for use and the respective deltas to be provided through the navigation message(s). UTC may be one, and a selected GNSS the other.

Figure 9.19 shows an example measurement of the difference between broadcast GGTO and a posteriori measured GGTO, in August/September 2015.

During the IOV phase, the broadcast GGTO parameters were computed based on time transfer between the Galileo Precise Time Facility and the US Naval Obser-

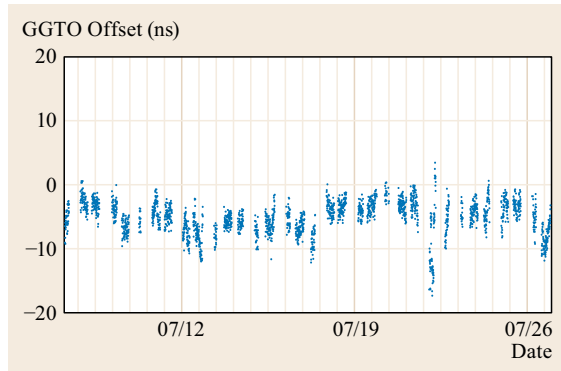


Fig. 9.19 Deviation of broadcast GGTO from measured GGTO in the period 25 November – 6 December 2013

vatory. In the completed Galileo ground segment the GGTO is being measured using calibrated Galileo-GPS receivers, which will improve the accuracy of this parameter.

9.3 Spacecraft

Galileo satellites are 700 kg/1500 W class spacecraft (Fig. 9.20, [9.32]). The first Galileo satellites GSAT0101...0104 were manufactured by EADS Astrium GmbH as satellite prime, and launched in pairs: a first dual launch to orbital plane A on 21 October 2011, followed by a dual launch to plane B on 12 October 2012.

The next family of Galileo satellites GSAT02xx is being manufactured by OHB System AG in Bremen, Germany, as prime [9.33]. A total of 22 satellites will be produced under this order. The first dual launch of this family in August 2014 was impaired by a malfunction

in the Fregat stage of the Soyuz launcher, and the satellites GSAT0201 and -0202 were placed into a nonnominal elliptic orbit. In 2015 their orbits were corrected as far as possible using the onboard resources. The navigation payloads were then activated and tested, to validate technology and performance, and the satellites since then broadcast navigation signals and are used for clock technology validation. Since nominal orbits were not reached these satellites cannot be accommodated in parts of the navigation message. GSAT0201 and -0202 are however healthy, and ESA is working to incorporate these satellites into the ground processing. The follow-

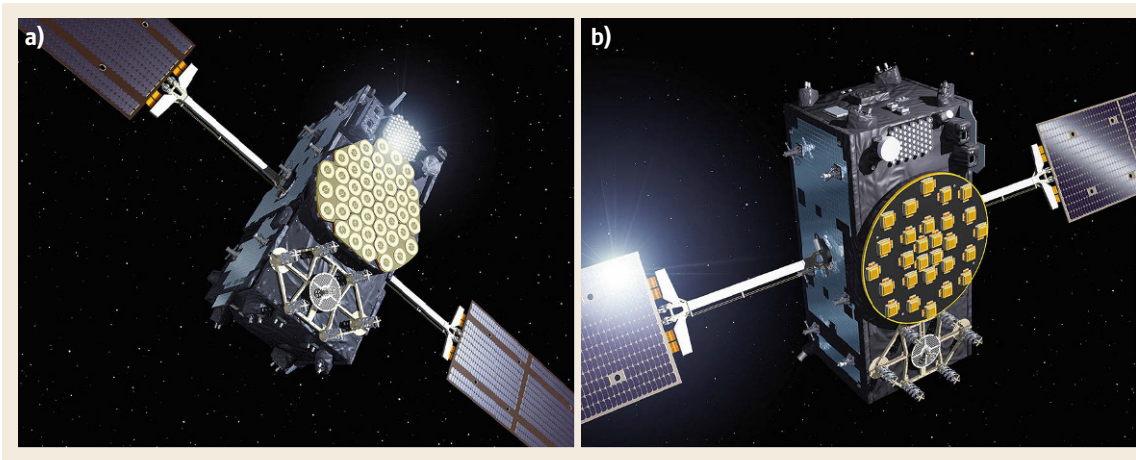


Fig. 9.20a,b Galileo satellites (artistic pictures, (a) GSAT010x, (b) GSAT02xx), (courtesy of ESA/P. Carril)

ing dual launches in March 2015, September 2015 and December 2015 reached the foreseen orbits, and the satellites became part of the nominal constellation.

9.3.1 Satellite Platform

A Galileo satellite can be partitioned into the classical components of a satellite design, that is platform (Table 9.8) and payload. The platform then comprises further subsystems for onboard data handling and control, attitude and orbit control including propulsion, power generation and distribution, thermal control, telemetry, and a laser retro reflector.

The attitude and orbit control system (AOCS) of the Galileo satellites is using three-axis attitude control during all phases and maneuvers [9.34]. Several operational modes are derived to support the mission sequence of events:

1. During the launch and early orbit phase (LEOP) as well as in contingency situations and safe modes, dedicated acquisition modes are used for Earth or Sun acquisition.
2. Orbit acquisition, station-keeping maneuvers, and end-of-life (EOL) decommissioning can use a dedicated orbit change mode. According to the Galileo orbit design very limited need of station-keeping maneuvers is anticipated [9.1, 6].
3. Normal mode is the nominal operational mode with full nadir pointing performance. This mode uses yaw steering to orient the solar panels towards the Sun, and to support the thermal control of the satellite.

In normal mode, the AOCS sensor/actuator configuration is based on Earth and Sun sensors for keeping the satellite pointed at Earth: the infrared Earth sensors detect the temperature contrast between the cold of deep space and the warmth of Earth's atmosphere, while the Sun sensors are visible-light detectors measuring the angle to the Sun. Angular momentum provided through reaction wheels is used to control attitude and rotational rate. The satellite rotates twice per orbit, to facilitate Sun-pointing of the solar wings. The angular momentum gradually accumulated by the reaction wheels is unloaded through magnetorquers, magnetic coils controlled to deliver the needed torque through interaction with the Earth magnetic field. Gyros are available for additional rate sensing. In operational modes except the nominal mode, thrusters can be used for impulse and attitude control.

The propulsion subsystem is based on monopropellant thrusters. The propulsion subsystem is typically equipped with a set of eight thrusters. Each thruster provides, under beginning of life (BOL) conditions,

a nominal thrust of 1 N using monopropellant-grade hydrazine. GSAT010x and -02xx are designed for direct injection into the final orbit, thus their propulsion subsystems need to provide only the delta-velocity capability needed for orbit correction maneuvers.

The power subsystem is responsible for the generation, storage conditioning and distribution of relevant power to the satellite. For the GSAT010x and -02xx families a classical regulated 50V bus architecture has been selected, which consists of

1. A power conditioning and distribution unit providing electrical power to all units on board the spacecraft
2. Two solar array wings supplying electrical power to the spacecraft during sun exposition and in parallel charging the battery after the eclipse phases
3. And a Li-Ion battery, storing the power provided by the solar arrays during the Sun phases and providing it during the eclipse phases.

The TT&C subsystem of GSAT010x and -02xx satellites is the link to the ground control segment by providing redundant command reception and telemetry transmission at S-Band. The TT&C subsystem operates in both ESA standard TT&C mode and spread spectrum mode. Accurate range-rate (Doppler) measurements are possible when the S-Band transponder is operated in coherent mode. S-Band TT&C operations are provided via hemispherical-coverage helix antennas situated on opposite sides of the satellite. Designed for orthogonal circular polarization, together they provide omnidirectional coverage for reception and transmission. Ranging operation is performed simultaneously with telemetry transmission.

The laser retro reflector (LRR) allows the measurement of the satellite's distance to within a few centimeters by reflecting a laser beam emitted from a ground station. The cat's eye reflector array of the LRR can be recognized on the Nadir panels of both Galileo satellite families shown in Fig. 9.20, just aside the navigation transmit antenna of the satellite. Laser ranging campaigns using the LRR are planned to be performed on average about once a year. In between the LRR campaigns, altitude measurements via S-band telemetry and telecommand link are used, which are sufficiently accurate to serve as intermediate measures.

9.3.2 Satellite Payload Description

The payload of Galileo satellites comprises a fully redundant triple-band navigation payload, and a SAR repeater [9.35].

The navigation payload can be functionally grouped into the mission uplink data handling system, a tim-

Table 9.8 Overview of main satellite platform characteristics

	GSAT010x	GSAT02xx
Satellites	4	22
Mass at launch	Approx. 700 kg	Approx. 715 kg
Size with solar array deployed	2.7 m × 1.6 m × 14.5 m	2.5 m × 1.1 m × 14.7 m
Design lifetime	12 years	12 years
Available power	1.4 kW	1.9 kW

ing subsystem, and the signal generation and transmitter subsystem.

The mission uplink data handling system receives the navigation message data and all related support data, through a dedicated code division multiple access (CDMA)-type C-band uplink served by the uplink stations of the Galileo ground segment.

The timing subsystem generates the onboard frequency reference, derived from an atomic clock as reference. Two different types of onboard clock technology are deployed, that is a rubidium atomic frequency standard and a passive hydrogen maser as shown in Fig. 9.21. The navigation payload contains two units of each technology, four clocks in total. Nominally one clock is operated as master clock, and one clock is hot redundant spare. The interface between the four clocks and the navigation signal generator unit is provided by a dedicated clock monitoring and control unit, which is also used for synchronization of master clock and active spare clock [9.36, 37]. This allows the spare to take over seamlessly should the master clock fail.

The navigation payload provides navigation signals on three carriers E1, E6 and E5 in the lower L-band. The generation of all navigation signals and their components is strictly coherent to the common frequency

reference from the timing subsystem. E5 offers a wide-band AltBOC signal of more than 50 MHz bandwidth containing two BPSK(10) subcarriers 30.69 MHz apart, each providing a pilot/data pair. The open signals on E6 and E1 are BPSK- and BOC-type modulations offering 31–41 MHz usable bandwidth (Table 9.4). The navigation signals are radiated through a dual-band transmit antenna, using a common antenna subsystem for E5 and E6 [9.38, 39].

Stability of the onboard reference frequency is one of the core performance parameters for the quality of the navigation payload. Example Allan deviation (ADEV) measurements from mid-2015 are shown in Fig. 9.22, covering exemplary RAFS and PHM results of GSAT010x and GSAT02xx satellites. The typical performance of the two clock technologies is clearly discernible.

The search-and-rescue repeater provides enhanced distress localization functionality for the provision of a SAR service. It is part of the Cospas-Sarsat MEOSAR System, and is interoperable with other MEOSAR repeaters on Globalnaja Navigazionnaja Sputnikowaja Sistema (Russian Global Navigation Satellite System) (GLONASS) and future GPS satellites [9.8–10]. The SAR transponder on Galileo satellites receives distress alerts in the 406.0–406.1 MHz band from any Cospas-

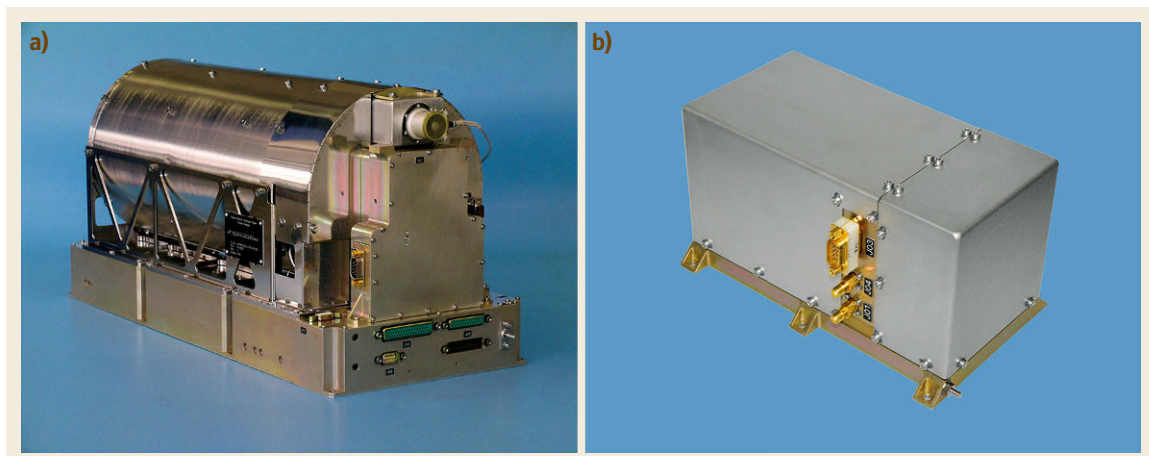


Fig. 9.21a,b Galileo passive hydrogen maser clock (a) and rubidium atomic frequency standard (b) (courtesy of Spectratime)

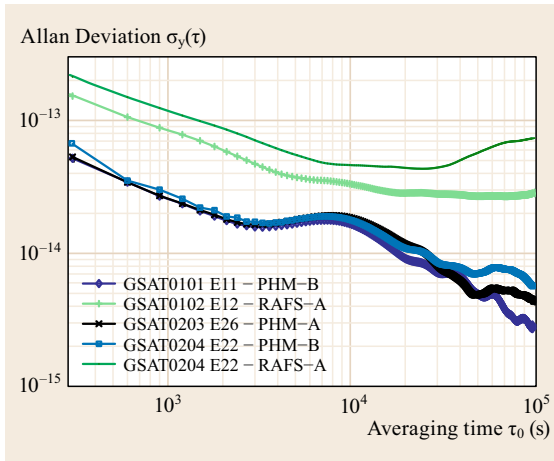


Fig. 9.22 Galileo PHM and RAFS frequency stability in mid-2015

Sarsat beacon, translates them to the SAR down-link band at 1544 MHz and rebroadcasts this signal to dedicated ground stations, medium altitude Earth orbit (MEO) Local User Terminals, which perform beacon localization in near real time based on difference of arrival (DOS) measurements of time and frequency [9.40].

Galileo also provides a SAR return link service, initially foreseen to inform the alerting beacon and thus the distressed people that the distress message has been received by the Cospas-Sarsat system. This acknowledgment is embedded in the navigation message [9.41].

9.3.3 Launch Vehicles

Galileo satellites are launched from the Guiana Space Centre, Europe's Spaceport in French Guiana, using Soyuz and Ariane launchers. The first twelve satellites were using Soyuz launchers, in a series of dual launches on October 2011, October 2012 (Fig. 9.23), August 2014, and March, September and December 2015.

The Soyuz launcher is the workhorse of the Russian space program, in continuous production since the 1960s, and a descendant in design terms of the R-7 rocket that launched Sputnik 1 in 1957, inaugurating the Space Age.

Soyuz has performed more than 1700 manned and unmanned missions. It is designed to extremely high reliability levels for its use in manned missions – today supporting operations of the International Space Station. The launch of GSAT0201 and -0202 in August 2014 ending in a nonnominal orbit will therefore hopefully remain an exception.

For French Guiana launches, this three-stage rocket plus the Fregat upper stage is assembled horizontally in the traditional Russian approach, then moved to the vertical so that its payload can be mated from above in the standard European way. A new mobile launch gantry aids this process, while also protecting the satellites and the launcher from the humid tropical environment.

For Galileo, a specially designed dispenser holds the two IOV satellites in place side by side during launch and then releases them sideways into their final orbits.



Fig. 9.23 Soyuz Launch Base at the European Space Port (CSG) French Guiana, ©ESA/S. Corvaja

A special version of the Soyuz launcher is also being used: the more powerful Soyuz ST-B variant, including a Fregat-MT upper stage, delivers the Galileo satellites into their final circular 23 222 km orbit.

The re-ignitable Fregat was previously used in its baseline version to deliver ESA's GIOVE-A and -B experimental satellites. Fregat-MT carries an additional 900 kg of propellant.

Alternatively a requalified Ariane 5 ES *Galileo* is available, able to deploy four Galileo satellites into

MEO orbit. The first such fourfold launch is foreseen for the last quarter of 2016.

The Ariane 5 ES version is an evolution of the initial Ariane 5 generic launcher that has been upgraded to allow re-ignition and long coast phases. These capabilities are necessary to inject a cluster of four Galileo satellites into their operational orbit. Re-ignition is also required to vacate the injection orbit after releasing the payload, for graveyarding of the last launcher stage outside the nominal Galileo orbit.

9.4 Ground Segment

The Galileo ground segment (Fig. 9.24) comprises a Ground Control Segment (GCS) for satellite and constellation control, and a Ground Mission Segment (GMS) for service-related tasks.

The ground control segment performs all functions related to command and control of the satellite constellation. It includes a worldwide network of S-band

TT&C stations hosted on Galileo remote sites, to provide global coverage.

The ground mission segment measures and monitors the Galileo navigation signals, computes the navigation message data and distributes it to the satellites. For this purpose the GMS includes two worldwide networks of stations:

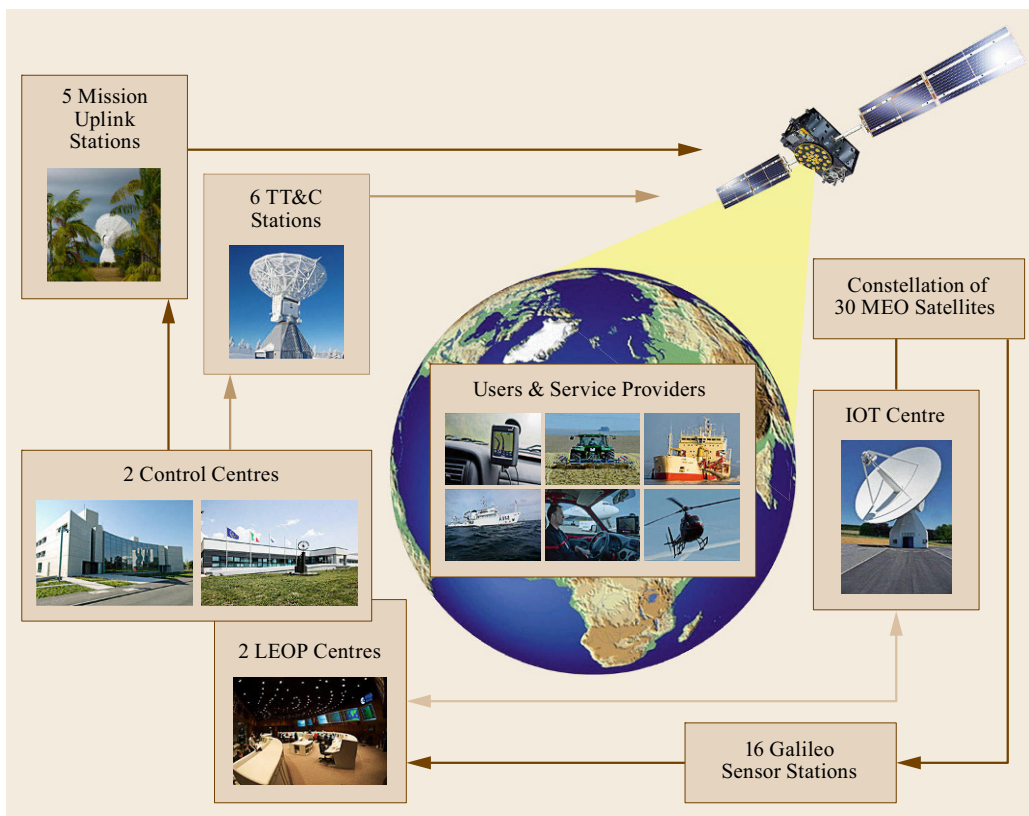


Fig. 9.24 Galileo system overview (courtesy of ESA/M. Pedoussaut, ESA/S. Corvaja, ESA/J. Mai, ESA/J. Huart, DLR, Telespazio)

1. L-band Galileo sensor stations (GSS) to collect ranging measurements of the Galileo navigation signals, for orbit determination, time synchronization and monitoring of the signal in space
2. C-band uplink stations (ULS) to uplink mission data (e.g., ephemerides and clock prediction, SAR return link and commercial service data).

GCS and the GMS core facilities are deployed in two Galileo control centers (GCC) located in Oberpfaffenhofen (Germany) and Fucino (Italy). A global data dissemination network connects all ground facilities. In their final configuration both control centers and the data dissemination network will be redundant such as to ensure service and operations continuity.

Two additional launch and early operation (LEOP) control centers (LOCC) located in Toulouse (French Space Agency, CNES) and Darmstadt (ESA Operations Centre, ESOC) provide the necessary services to take control of the satellites after their separation from the launch vehicle and until they have reached their position within the assigned orbital slots.

Each LEOP is followed by in-orbit testing (IOT) to verify satellite payload health and survival of the launch. These tests are supported by the designated IOT station in Redu (Belgium), which comprises a calibrated high-gain antenna and measurement system for the L-band navigation signals, and testing equipment and transmitters for the C-band and SAR UHF RF links. A second use of the IOT station Redu is for regular signal characterization during routine operations.

The ground segment manages interfaces to the satellite manufacturers, needed for onboard software maintenance, operations support and telemetry analysis, and in support of eventual troubleshooting of satellite platform and payload units.

Further external interfaces of the Galileo ground segment are installed, to connect to entities contributing to the provision of Galileo services:

1. The GNSS service center (GSC), foreseen as an interface between the Galileo system and external data providers for the Galileo open service (OS) and the Galileo commercial service (CS). The GSC facility is located near Madrid, Spain.
2. The Galileo security monitoring centers (GSMC) providing system security monitoring, management of the public regulated service user segment, and the point of contact platforms (POCP) to interface with national competent PRS authorities (CPA). GSMC facilities are located in France and in the United Kingdom.
3. The time and geodetic reference service providers (TSP, GRSP) to monitor and steer Galileo System Time and Galileo Terrestrial Reference Frame vis-a-vis international meteorological standards.
4. The SAR Galileo data service provider (SAR GDSP) to carry out the position determination of the distress alert emitting beacons once they have been detected by the dedicated ground segment, and to provide SAR return link messages for dissemination through the Galileo navigation signals. The SAR GDSP premises are located in Toulouse, France.
5. The Galileo reference center (GRC) to provide independent performance monitoring of the Galileo services. The GRC facility will be located in Noordwijk, the Netherlands.

These service providers are procured, coordinated and operated by the European GNSS Agency (GSA), an institution of the European Union tasked with the operation of the Galileo system, and in charge of service provision and quality of the operational system.

9.5 Summary

Galileo is a joint initiative of ESA and the European Commission (EC), to deploy a highly accurate and independent GNSS under civilian control. Compatibility and interoperability with existing GNSSs and especially with GPS were important requirements during the concept and design studies. The procurement of the Galileo system was launched in 2008, and has since then proceeded with space and ground segment development, manufacturing and deployment. The number of signals from Galileo satellites available for testing and development is increasing with deployment, and this enabled the first Galileo-only position fix to begin in 2013 and suc-

cessful in-orbit validation in the second half of 2013. The simultaneous successful build up of the ground mission and control segments with their worldwide infrastructure is bringing the system into an operational state. Interfaces to external service providers as well as the services themselves are being installed. All these deployments are accompanied by a continuous process of system verification and tuning, visible in the steady improvement of the already good initial results. The sum of these efforts will allow the start of initial services in 2016, with nine nominal satellites plus two satellites in nonnominal orbits, and with the prospect of six more

satellites to be launched in 2016. Successful validation of this system configuration will be the starting point of the exploration phase, planned for 2017, where Galileo will become officially available. The full constellation of 24 satellites plus in-orbit spares and the completion of the ground segment will be reached in 2020.

Acknowledgments. The authors would like to thank their colleagues in the Galileo Project Office and all teams at Industry contributing to the European project for the Global Navigation Satellite System Galileo. Without their continuous efforts Galileo would not have come into existence.

References

- 9.1 R. Zandbergen, S. Dinwiddy, J. Hahn, E. Breeuwer, D. Blonski: Galileo orbit selection, ION GNSS 2004, Long Beach (ION, Virginia 2004) pp. 616–623
- 9.2 R. Píriz, B. Martín-Peiró, M. Romay-Merino: The Galileo constellation design: A systematic approach, ION GNSS 2005, Long Beach (ION, Virginia 2005) pp. 1296–1306
- 9.3 D. Blonski: Galileo IOV and first results, Proc. ENC 2013, Vienna (Austrian Inst. of Navigation, Vienna 2013)
- 9.4 D. Blonski: Performance Extrapolation to FOC and outlook to Galileo early services, Proc. ENC 2014, Rotterdam (Netherlands Institute of Navigation, Rotterdam 2014)
- 9.5 D.A. Vallado: *Fundamentals of Astrodynamics and Applications (Space Technology Library)*, 4th edn. (Microcosm, California 2013)
- 9.6 D. Navarro-Reyes, A. Notarantonio, G. Taini: Galileo constellation: Evaluation of station keeping strategies, 21st Int. Symp. Space Flight Dynamics (ISSFD), Toulouse (CNES, Toulouse 2009)
- 9.7 R.B. Langley: Dilution of precision, GPS World **10**(5), 52–59 (1999)
- 9.8 Specification for Cospas-Sarsat 406 MHz Distress Beacons, C/S_T.001, Issue 3, Revision 16, December 2015, Cospas-Sarsat
- 9.9 Description of the 406 MHz Payload Used in the Cospas-Sarsat MEOSAR System, C/S_T.016, Issue 1, Revision 1, December 2015, Cospas-Sarsat
- 9.10 Cospas-Sarsat MEOLUT Performance Specification and Design Guidelines, C/S_T.019, Issue 1, December 2015, Cospas-Sarsat
- 9.11 European GNSS (Galileo) Open Service Signal In Space Interface Control Document, OS SIS ICD, Iss. 1.2, Nov. 2015 (European Union 2015)
- 9.12 L. Ries, J.-L. Issler, O. Julien, C. Macabiau: Method of Reception and Receiver For a Radio Navigation Signal Modulated by a CBOC Spread Wave Form, Patents US8094071, EP2030039A1 (Centre National d'Études Spatiales 2012)
- 9.13 O. Julien, C. Macabiau, L. Ries, J.-L. Issler: 1-Bit processing of composite BOC (CBOC) signals and extension to time-multiplexed BOC (TMBOC) signals, ION NTM 2007, San Diego (ION, Virginia 2007) pp. 227–239
- 9.14 A. De Latour, G. Artaud, L. Ries, F. Legrand, M. Sihrener: New BPSK, BOC and MBOC tracking structures, ION ITM, Anaheim (ION, Virginia 2009) pp. 396–405
- 9.15 L. Ries, L. Legrand, L. Lestarquit, W. Vigneau, J.-L. Issler: Tracking and multipath performance assessments of BOC signals using a bit level signal processing simulator, Proc. ION GPS 2003, Portland (ION, Virginia 2003) pp. 1996–2010
- 9.16 M. Soellner, P. Erhard: Comparison of AWGN tracking accuracy for alternative-BOC, complex-LOC and complex-BOC modulation options in Galileo E5 band, Proc. ENC GNSS 2003, Graz (Austrian Institute of Navigation, Graz 2003)
- 9.17 L. Lestarquit, G. Artaud, J.-L. Issler: AltBOC for dummies or everything you always wanted to know about AltBOC, Proc. ION GNSS, Savannah (ION, Virginia 2008) pp. 961–970
- 9.18 European GNSS (Galileo) Open Service Ionospheric Correction Algorithm for Galileo Single Frequency Users, Iss. 1.2, Sep. 2016 (European Commission 2016)
- 9.19 I. Fernández-Hernández, I. Rodríguez, G. Tobías, E. Carbonell, G. Seco-Granados, J. Simón, R. Blasi: The Galileo commercial service: Current status and prospects, Inside GNSS **10**(1), 38–48 (2015)
- 9.20 J. Saastamoinen: Atmospheric correction for the troposphere and the stratosphere in radio ranging satellites. In: *The Use of Artificial Satellites for Geodesy*, Geophys. Monogr., Vol. 15, ed. by S.W. Henriksen, A. Mancini, B.H. Chovitz (AGU, Washington 1972) pp. 247–251
- 9.21 R. Orus-Perez, R. Prieto-Cerdeira, B. Arbesser-Rastburg: The Galileo single-frequency ionospheric correction and positioning observed near the solar cycle 24 maximum, 4th Int. Colloq. Sci. Fundam. Asp. Galileo Program., Prague (ESA, Noordwijk 2013)
- 9.22 R. Prieto-Cerdeira, S. Binda, M. Crisci, I. Hidalgo, D. Rodriguez, V. Borrel, J. Giraud: Ionospheric propagation activities during GIOVE Mission experimentation, 4th Eur. Conf. Antennas Propag. (EuCAP), Barcelona (IEEE, 2010) pp. 1–6
- 9.23 A. Martellucci, R. Prieto-Cerdeira: Review of tropospheric, ionospheric and multipath data and models for Global Navigation Satellite Systems, 3rd Eur. Conf. Antennas Propag. (EuCAP), Berlin (IEEE, 2009)
- 9.24 E. Breeuwer, S. Binda, G. Lopez-Risueno, D. Blonski, F. Gonzalez Martinez, A. Mudrak, R. Prieto-Cerdeira, I. Stojkovic, J. Hahn, M. Falcone: Galileo works, Inside GNSS **8**(2), 60–66 (2013)
- 9.25 R. Prieto-Cerdeira, R. Orus-Perez, E. Breeuwer, R. Lucas-Rodriguez, M. Falcone: Performance of the Galileo single-frequency ionospheric correction during in-orbit validation, GPS World **25**(6), 53–58 (2014)

- 9.26 M. Falcone, S. Binda, E. Breeuwer, J. Hahn, E. Spinelli, F. Gonzalez, G. Lopez Risueno, P. Giordano, R. Swinden, G. Galluzzo, A. Hedquist: Galileo on its own: First position fix, *Inside GNSS* **8**(2), 50–71 (2013)
- 9.27 S. Binda: Galileo timing performance update, plenary presentation, Proc. ENC 2014, Rotterdam (Netherlands Institute of Navigation, Rotterdam 2014)
- 9.28 Standard-Frequency and Time-Signal Emissions, ITU-R Recommendation TF.460-6 (International Telecommunication Union, Radio-communication Bureau, Geneva, 2002) pp. 460–466
- 9.29 R. Píriz, Á.M. García, G. Tobías, V. Fernández, P. Tavella, I. Sesia, G. Cerretto, J. Hahn: GNSS interoperability: Offset between reference time scales and timing biases, *Metrologia* **45**(6), 87–102 (2008)
- 9.30 P. Defraigne, W. Aerts, G. Cerretto, G. Signorile, E. Cantoni, I. Sesia, P. Tavella, A. Cernigliaro, A. Samperi, J.M. Sleewaegen: Advances on the use of Galileo signals in time metrology: Calibrated time transfer and estimation of UTC and GGTO using a combined commercial GPS-Galileo receiver, Proc. 45th PTI Syst. Appl. Meet., Bellevue (ION, Virginia 2013) pp. 256–262
- 9.31 J.H. Hahn, E.D. Powers: Implementation of the GPS to Galileo time offset (GGTO), IEEE Int. Freq. Control Symp. Expo., Vancouver (IEEE, 2005)
- 9.32 Galileo navigation program: FOC (Full Operational Capability), eoPortal, ESA, 2016, <https://directory.eoportal.org/web/eoportal/satellite-missions/g/galileo-foc>
- 9.33 K. Pauly: Galileo FOC – Design, production, early operations after 1st launch, and project status, IAC-14-B2.2.1, 65th Int. Astronaut. Cong. (IAC), Toronto (IAF, Paris, France 2014) pp. 1–4
- 9.34 A. Konrad, H.-D. Fischer, C. Müller, W. Oesterlin: Attitude and orbit control system for Galileo IOV, Proc. 17th IFAC Symp. Autom. Control Aerosp., Toulouse, ed. by H. Siguerdidjane (IFAC, Laxenburg, Austria 2007) pp. 25–30
- 9.35 G.T.A. Burbidge: Development of the navigation payload for the Galileo in-orbit validation (IOV) phase, IGNS Symp. 2007, Sydney (IGNS Society, Tweed Heads 2007)
- 9.36 D. Felbach, D. Heimbuerger, P. Herre, P. Rastetter: Galileo payload 10.23 MHz master clock generation with a clock monitoring and control unit (CMCU), Proc. IEEE FCS and 17th EFTF 2003, Tampa (IEEE, 2003) pp. 583–586 doi:10.1109/FREQ.2003.1275156
- 9.37 D. Felbach, F. Soualle, D. L. Stopfkuchen, A. Zeninger: Clock monitoring and control units for navigation satellites, IEEE FCS 2010, Newport Beach (IEEE, 2010) pp. 474–479 doi:10.1109/FREQ.2010.5556283
- 9.38 A. Montesano, C. Montesano, R. Caballero, M. Naranjo, F. Monjas, L.E. Cuesta, P. Zorrilla, L. Martinez: Galileo system navigation antenna for global positioning, Proc. 2nd EuCAP, Edinburgh (IET, Stevenage 2007) pp. 1–6
- 9.39 P. Valle, A. Netti, M. Zolesi, R. Mizzoni, M. Bandinelli, R. Guidi: Efficient dual-band planar array suitable to Galileo, Proc. 1st EUCAP, Nice (IEEE, 2006) pp. 1–7 doi:10.1109/EUCAP.2006.4584868
- 9.40 F. Paggi, I. Stojkovic, D. Oskam, E. Breeuwer, M. Gotta, M. Marinelli: SAR/Galileo IOV forward link test campaign results, Proc. ENC-GNSS 2014, Rotterdam (Netherlands Institute of Navigation, Rotterdam 2014)
- 9.41 F. Paggi, I. Stojkovic, A. Postinghel, D. Ratto, E. Breeuwer, M. Gotta: SAR/Galileo IOV return link test campaign results, Proc. ENC-GNSS 2014, Rotterdam (Netherlands Institute of Navigation, Rotterdam 2014)

Chinese Navi

10. Chinese Navigation Satellite Systems

Yuanxi Yang, Jing Tang, Oliver Montenbruck

This chapter introduces the BeiDou (COMPASS) Navigation Satellite System from its early stage as a demonstration system to its evolution to a global system. First, the development strategy and basic principle of BeiDou Demonstration System are reviewed. Its basic performance is given in details. Second, the basic information of BeiDou (regional) system including constellation, frequency, coordinate reference system, and time datum is described. Its initial performance is evaluated by using single-point positioning, code and carrier phase differential positioning. Some application examples are introduced. Third, the BeiDou (Global) Navigation Satellite System (BDS) is described. Position dilution of precision is analyzed and BeiDou's contribution is summarized. At last, Chinese Area Positioning System is briefly introduced.

10.1	BeiDou Navigation Satellite Demonstration System (BDS-1)	275
10.1.1	System Architecture and Basic Characteristics	275
10.1.2	Navigation Principle	277
10.1.3	Orbit Determination	278
10.1.4	Timing	278

10.2	BeiDou (Regional) Navigation Satellite System (BDS-2)	279
10.2.1	Constellation	279
10.2.2	Signals and Services	281
10.2.3	Navigation Message	283
10.2.4	Space Segment	286
10.2.5	Operational Control System	288
10.2.6	BeiDou Satellite-Based Augmentation System	289
10.2.7	Coordinate Reference System	290
10.2.8	Time System	291
10.3	Performance of BDS-2	293
10.3.1	Service Region	293
10.3.2	Performance of Satellite Clocks	293
10.3.3	Positioning Performance	295
10.3.4	Application Examples	297
10.4	BeiDou (Global) Navigation Satellite System	297
10.5	Brief Introduction of CAPS	298
10.5.1	CAPS Concept and System Architecture	298
10.5.2	Positioning Principle of CAPS	300
10.5.3	Trial CAPS System	301
	References	301

The Big Dipper, the Plough or the Saptarishi (in Chinese 北斗, BěiDǒu) are well recognized in many cultures as the most important set of stars giving directions to people in the Northern Hemisphere. Besides telling the north direction, the Big Dipper also indicates the seasons. An ancient Chinese book already states that it is spring when the dipper handle directs east, summer when it directs down, autumn when it directs west, and winter when it directs up.

The Big Dipper can only be used for orientation in clear nights. As the first manmade navigation device, which was not affected by weather and could always identify the cardinal directions during day and night, the ancient compass was invented in China [10.1]. It

is composed of a box frame and magnetic needle, or a spoon made from lodestone (Fig. 10.1). This magnetic device is used as a means of orientation, which always points in the northern (or southern) direction.

In Chinese mythology, the army of Huangdi (the *Yellow Emperor*, who was regarded as the initiator of Chinese civilization) lost their direction in a heavy fog around 2697 BC. In a dream, a Fairy clued to Huangdi that the compass could show the south direction. Then the compass was invented. The compass vehicle used in the war helped Huangdi to win the war against Chi You (who was a tribal leader of ancient China). The above story is just one of the many beautiful legends left to Chinese. More than that, the compass was widely



Fig. 10.1 A Chinese ancient compass (called *Sinan*) made up of a spoon from loadstone and a bronze plate with direction markings. Reproduced with permission of Panorama Media, Inc.

used in ancient China from the Qin Dynasty to Early Ming dynasty. It helped the traders and sailors find the direction during their voyages and expeditions around Southeast Asia and Pacific and Indian Ocean [10.3].

As Chinese scientists and engineers started to implement their own navigation satellite system, *BeiDou*

was given as its name which was naturally chosen and recognized by the Chinese people. At the same time, its English name *COMPASS* has been also used for a dozen of years, which is largely used in the system’s official frequency filing to the International Telecommunication Union (ITU).

China decided to build an independent navigation satellite system in the 1980s. Three steps for constructing the system were planned (Fig. 10.2). The first step is the BeiDou Navigation Satellite Demonstration System which is called BeiDou-1 or simply BDS-1 [10.2]. In 1994, China started the buildup of the BeiDou Navigation Satellite Demonstration System. In 2000, two BeiDou navigation experiment satellites were launched. In 2003, the third BeiDou satellite was launched, further enhancing the system’s performance. Thus, BDS-1 was formally established, which made China the third nation in possession of an independent navigation satellite system following the United States and Russia.

The second step is the regional BeiDou Navigation Satellite System. Implementation of BDS-2 was started in 2004, and the first satellite, a medium Earth orbit (MEO) satellite was launched in 2007. An op-

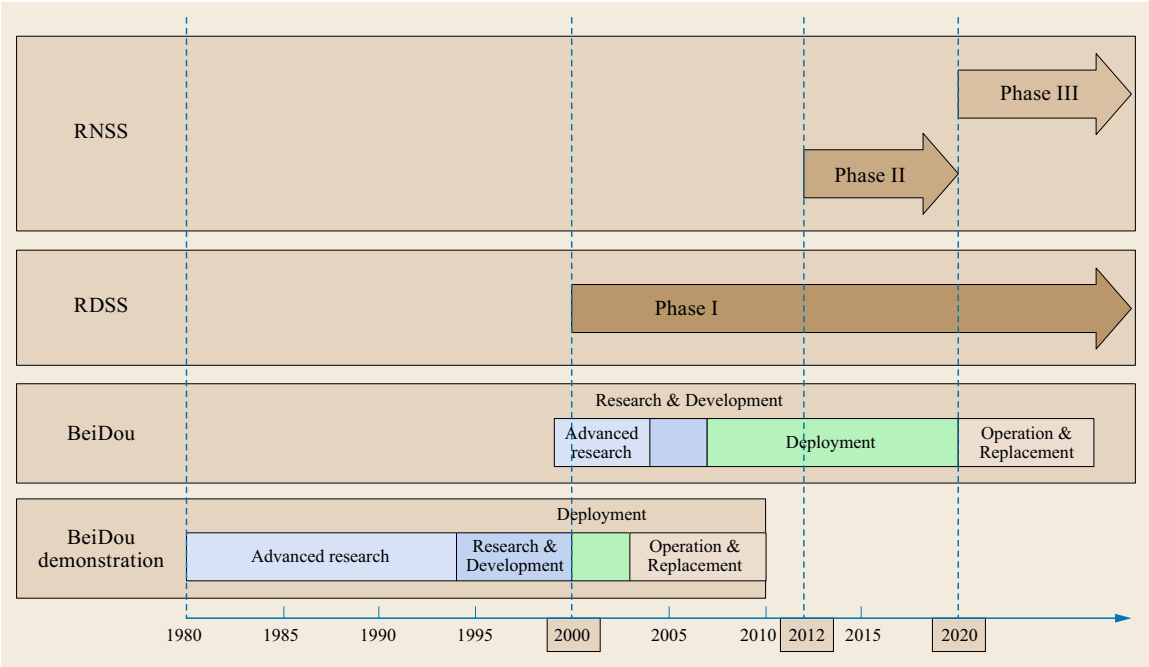


Fig. 10.2 The three phases in the development of the BeiDou Navigation Satellite System. Following the implementation of a regional radio determination satellite service (RDSS) in the frame of BDS-1, a regional radio navigation satellite service (RNSS) was established by BDS-2 in 2012. This will ultimately be extended to a global service by the third generation BeiDou Navigation Satellite System, BDS-3. After [10.2], reproduced with permission of Beijing Satellite Navigation Center

erational navigation service available for China and large parts of Asia-Pacific region was declared by the end of 2012 [10.4]. It is accomplished through a constellation of 14 satellites, including five satellites in geostationary Earth orbit (GEO), five satellites in in-

clined geosynchronous orbit (IGSO), and four MEO satellites.

As a third step, the BeiDou Navigation Satellite System with global coverage (BDS-3) is built-up, which will be completed around 2020.

10.1 BeiDou Navigation Satellite Demonstration System (BDS-1)

The BeiDou Navigation Satellite Demonstration System (BDS-1; [10.5]) offers a combined localization and communication service through a pair of geostationary satellites. Its main functions comprise:

- Positioning (navigation): quick determination and provision of the user's location.
- Short message communication: provision of two-way message exchange between the users and the master control station (MCS), as well as among the users themselves.
- Timing: broadcast of timing information and provision of time delay corrections for timing users.

All of these functions are achieved through the same channel. The concept is known as *radio determination satellite service* (RDSS, [10.6]) and has first been proposed for the Geostar system. Geostar was developed in the United States throughout the 1980s as a civil localization and communication system [10.7], but soon abandoned in view of the emerging availability of GPS.

Fig. 10.3 GEO satellite of the BeiDou Navigation Satellite Demonstration System (BDS-1). Reproduced with permission of Beijing Satellite Navigation Center ►

Key differences between the BDS-1 RDSS and the RNSS of other constellations such as GPS, GLONASS, and BDS-2/3 are highlighted in Table 10.1.

10.1.1 System Architecture and Basic Characteristics

Similar to other satellite navigation systems, the BDS-1 system architecture comprises a space segment, a ground control segment, and the user terminals.

The initial BDS-1 constellation was made up of two geostationary satellites launched in late 2000 and positioned at 80°E and 140°E. According to their position, they are called BeiDou-West and BeiDou-East, simplified as BeiDou 1A and BeiDou 1B (Fig. 10.3). For backup purposes, a third satellite (BeiDou 1C) was added in 2003 at 110.5°E [10.5]. Meanwhile, all of

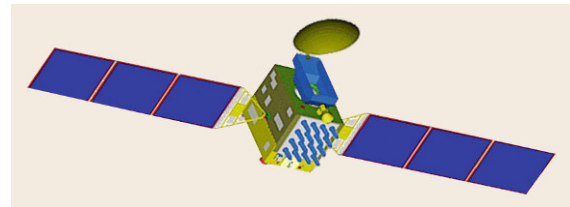


Table 10.1 Comparison of the BDS-1 RDSS with the RNSS of other satellite navigation systems

	BDS-1 RDSS	RNSS
Basic principles	Determination of the user's location by the master control station	Determination of the user's location and velocity by the user
Constellation	GEO	GEOs, MEOs, IGSOs
Service type	Positioning, timing, location report, short message communication	Positioning, timing, velocity determination
User transmits response signal	Yes	No
Observation	Sum pseudorange from the user to the master control station via the satellite	Pseudorange from satellite to user and Doppler measurement
Payload complexity	Low	High
Coverage	Regional	Regional or global coverage
Service frequency	Single service for low- and medium-dynamic users	Continuous service for low, medium and high dynamic users
Application	Positioning, location report, communication, rescue	Navigation

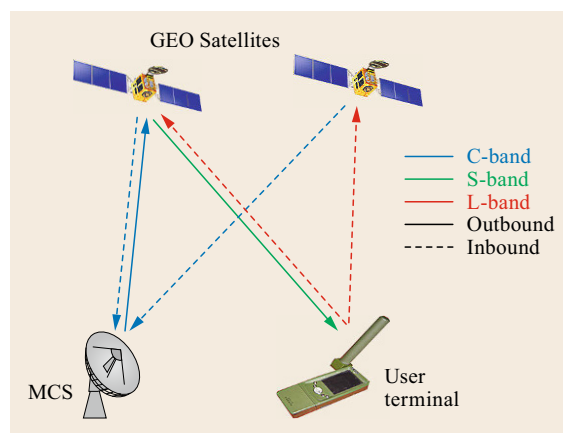


Fig. 10.4 RDSS links of BeiDou-1. Individual frequencies are distinguished by different colors. Solid lines refer to output signals (master control station to user), while dashed lines refer to inbound (return) signals

these spacecraft have reached their end of life and have been substituted by satellites of the second-generation BeiDou system. These hold the same positions in the geostationary belt and continue to provide a BDS-1-type RDSS in addition to their primary RNSS.

Each satellite has two outbound transponders and two inbound transponders. The outbound transponders transfer signals which are emitted from the MCS to the satellites and further to the users. Vice versa, the inbound transponders transfer the signals which are emitted from the user to the satellites and further to the MCS (Fig. 10.4). The feeder link, which includes the uplink from the MCS to the GEO satellite of outbound signals and the downlink of the inbound signal from the GEOs to the MCS, uses the C-band frequency allocation for Fixed Satellite Service. The service link, which includes the uplink of inbound signals from the user to the GEOs and the downlink of the outbound signal from the GEOs to the user, uses the L-band (1610–1626.5 MHz; uplink) and S-band (2483.5–2500 MHz; downlink) frequency allocation for RDSS.

The BDS-1 ground control segment consists of the MCS (Fig. 10.5), at Beijing and more than 20 Calibration Stations. The MCS is responsible for transmitting the outbound signals and for receiving the inbound signals, for performing satellite orbit determination and ionosphere correction, for determining the user location, and for sending the short message to the subscribed users. The calibration stations provide the basic measurements for orbit determination, wide area differential positioning, and user elevation computation from barometric altimeter data.

The information flow of BDS-1 is illustrated in Fig. 10.6. For a positioning request, the MCS de-



Fig. 10.5 BeiDou master control station at Beijing. Reproduced with permission of Beijing Satellite Navigation Center

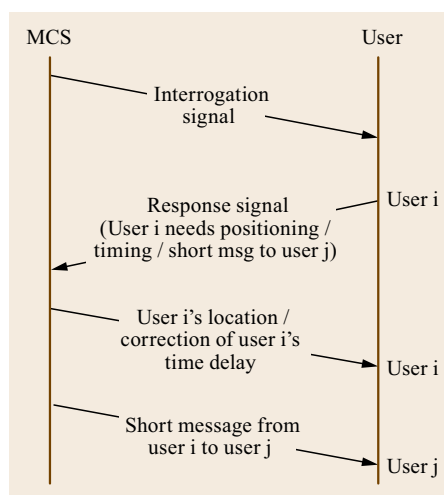


Fig. 10.6 Information flow of the BDS-1 radio determination satellite service

termines the user location from the measured signal turn-around time and the user's height. The latter is queried from the digital height database stored in the MCS or provided by the user. The resulting position information is then sent back to the user through the outbound signal. For a short message request, the MCS transmits the message to the addressee through the outbound signal. For timing requests, finally, the MCS calculates the precise correction of the user's time delay and sends it to the user through the outbound signal. The user then adjusts the local clock based on the time delay correction, thus, synchronizing it with the clock of the MCS.

The RDSS user terminals (Fig. 10.7) are capable of sending the requests and receiving location information and short messages. They can work in two modes. One mode is to receive one outbound signal and to transmit inbound signals via two satellites. Another mode is to receive the outbound signals from the two satellites

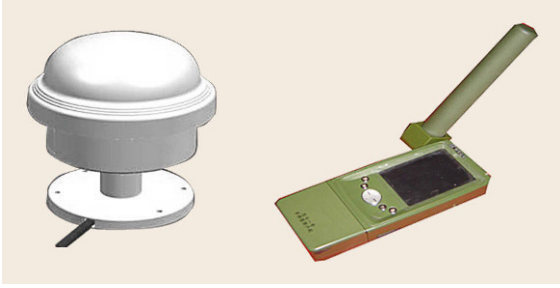


Fig. 10.7 BeiDou RDSS user terminals. Reproduced with permission of Beijing Satellite Navigation Center

when the user is located in the common coverage area of the two satellites and to transmit the inbound signal via only one satellite. According to the time difference of the two outbound signals received and the service request transmitted by one satellite, the MCS measures the pseudorange from the user to the two satellites and calculates the user position.

Following [10.5], the BDS-1 system is able to handle a total of 540 000 positioning requests per hour. To control the overall system utilization, users are divided into several service classes with update rates of 1–9 s, 10–60 s, and 60–120 s. While the BDS-1 RDSS is free of charges, potential users must register and obtain a card to uniquely identify their terminal for the localization and communication service.

The BDS-1 service area illustrated in Fig. 10.8 is governed by the location of its geostationary satellites as well as the calibration stations. Overall, it comprises China and its surrounding areas, from longitude 70°E to 140°E and from latitude 5°N to 55°N.

The BDS-1 service performance specification is summarized in Table 10.2. A positioning accuracy of about 20 m can be achieved if the terminal locates in the region with calibration stations; otherwise, the positioning accuracy is about 100 m. These specifications are in good accord with practical performance results reported in [10.5]. Here, horizontal positioning errors of about 8 m (2D, 1 σ) have been obtained both in static tests and a low dynamic test onboard a maritime vessel with BDS-1 receivers in comparison with known positions from a GPS reference receiver.

10.1.2 Navigation Principle

The localization of a user terminal in BDS-1 is based on turn-around signal travel time measurements initiated by the MCS. As mentioned before, the MCS first emits an interrogation signal to the two satellites, which is subsequently broadcast to the users in the service area via the outbound transponders of the two satellites. The user receives the interrogation signal and sends its re-

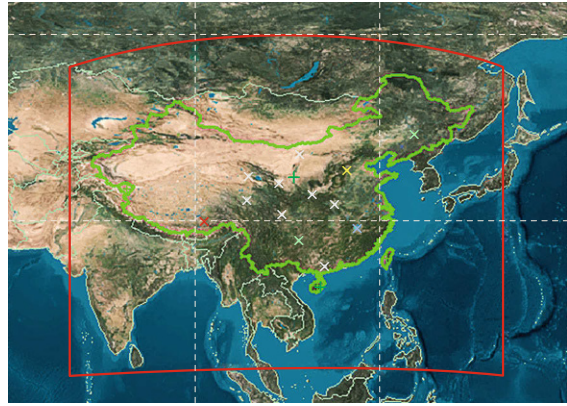


Fig. 10.8 Service coverage of BDS-1. Reproduced with permission of Beijing Satellite Navigation Center

Table 10.2 Performance specification for the BDS-1 radio determination satellite service (after [10.5])

Parameter	Value
Horizontal positioning accuracy	20 m (1 σ)
One-way time accuracy	100 ns
Two-way time accuracy	20 ns
Short message communication	120 Chinese chars./msg.

sponse signal with the user's service request back to the satellite (Figs. 10.6 and 10.4).

A first observation L_1 is obtained by measuring the difference between the transmit time t_t and the receive time $t_{r(s,s)}$ of a signal passing through the same satellite s on both the inbound and outbound links. Likewise, a second observation L_2 is obtained from the travel time $t_{r(s,s')} - t_t$ of a signal passing through satellite s on the outbound link but through the second satellite s' on the inbound link.

Denoting the MCS position by \mathbf{r}_m , the position of the two satellites by \mathbf{r}_s and $\mathbf{r}_{s'}$, respectively, and the (unknown) user position by $\mathbf{r}_u = (x_u, y_u, z_u)$, the two observations can be modeled as

$$\begin{aligned} L_1 &= c \cdot (t_{r(s,s)} - t_t) \\ &= 2\|\mathbf{r}_s - \mathbf{r}_u\| + 2\|\mathbf{r}_s - \mathbf{r}_m\| \end{aligned} \quad (10.1)$$

and

$$\begin{aligned} L_2 &= c \cdot (t_{r(s,s')} - t_t) \\ &= \|\mathbf{r}_s - \mathbf{r}_u\| + \|\mathbf{r}_{s'} - \mathbf{r}_u\| \\ &\quad + \|\mathbf{r}_s - \mathbf{r}_m\| + \|\mathbf{r}_{s'} - \mathbf{r}_m\|, \end{aligned} \quad (10.2)$$

respectively. These relations are likewise applicable for the operation mode in which the user receives signals from the two satellites and transmits them back via only one satellite.

It may be noted that the BDS-1 measurement model does not include explicit clock offset terms as found

in the traditional global navigation satellite system (GNSS) pseudorange model. This is due to the fact that the transmit and receive times are measured by a common clock in the MCS. The control segment (CS) clock offset from the BDS system time is common to the transmit and receive time stamps and cancels when forming the range measurement. As such, the minimum number of independent observations required for a position fix is one less than in other navigation satellite systems using one-way pseudorange observations. However, BDS-1 observations are still affected by equipment delays (transmitter, transponder, and receiver biases) which require proper consideration.

Since the positions of the satellites and the MCS are considered as known quantities, L_1 and L_2 are in fact equivalent to distance measurements of the user relative to the two satellites s and s' . These are not sufficient, though, to uniquely determine the three-dimensional user position. Information on the user's height h_u above the reference ellipsoid is, therefore, employed as an independent, third observation. h_u can either be provided from a barometric altimeter in the user terminal or provided from a digital elevation model maintained in the MCS for terrestrial users [10.5]. Given the ellipsoid height, a measurement

$$\begin{aligned} L_3 &= h_u + N \\ &= \sqrt{x_u^2 + y_u^2 + (z_u + Ne^2 \sin \varphi)^2} \end{aligned} \quad (10.3)$$

is formed, which describes the distance of the receiver from the Earth rotation axis along the normal line of the reference ellipsoid. Here,

$$N = \frac{a}{\sqrt{1 - e^2 \sin^2 \varphi}} \quad (10.4)$$

is the radius of curvature in the prime vertical at geodetic latitude φ , while a and e denote the semi-major axis and eccentricity of the reference ellipsoid (Sect. 2.2.1).

Equations (10.1)–(10.3) cannot be solved directly for the unknown user position, but can be linearized in the vicinity of an approximate a priori value $\mathbf{r}_{u,0}$. This yields a three-dimensional set of linear equations from which corrections $\Delta \mathbf{r}_u = \mathbf{r}_u - \mathbf{r}_{u,0}$ are obtained. This process is repeated in an iterative manner until the solution is obtained with the desired accuracy.

For simplicity, atmospheric propagation effects and equipment specific delays have been neglected in the above discussion, but are taken into account in the actual MCS processing. Among others, these corrections are based on a distributed set of calibration stations enabling wide area differential positioning.

10.1.3 Orbit Determination

Similar to other satellite navigation systems, positioning with BDS-1 depends on proper knowledge of the satellite positions. In BDS-1, the orbit determination relies on measurements of the calibration stations, which follow the same principle as described in Sect. 10.1.2 for the user positioning. Calibration stations for orbit determination respond to the same frame interrogation signal from the MCS at a certain sample interval. Therefore, the MCS can measure a group of distances related to the individual calibration stations for orbit determination of each satellite.

The basic observation model for satellite orbit determinations matches (10.1) and (10.2) when replacing the user position \mathbf{r}_u by the (known) position of a calibration station. However, there are some differences between satellite orbit determination and user positioning. While the number of measurements in the positioning equations of BDS-1 matches with that of the unknown position parameters, the number of measurements in the orbit determination is usually greater than that of the orbit parameters (typically six per satellite). Also, the orbits of both BDS-1 satellites should be determined simultaneously, since the measurement equations for each calibration station depend on the position of two satellites (s and s') at a time.

10.1.4 Timing

Apart from positioning, the BDS-1 RDSS also supports the synchronization of user terminals with the BeiDou system time (BDT) maintained at the MCS. BDT is a continuous time scale without leap seconds [10.8] that is realized by composite clocks with robust data fusion. BDT is aligned to the realization of Coordinated Universal Time UTC(NTSC) maintained by the Chinese National Time Service Center (NTSC). UTC(NTSC) itself is traced to the Coordinated Universal Time (UTC) by satellite common view (CV) links and the offset of BDT with respect to UTC is controlled within 30 ns.

BDS-1 provides two types of timing services: one-way and two-way timings with 100 ns and 20 ns timing accuracies, respectively. Both of these are based on a comparison of transmit and receive times measured relative to the local clock after accounting for the known geometric propagation time and possible correction atmospheric and equipment delays.

Each second, the MCS transmits a total of 32 frames of 31.25 ms duration aligned with the integer second of BDT. Denoting by Δ_1 the time difference between the reception of the n -th frame and the receiver's preceding 1-pulse-per-second (1-pps) epoch (Fig. 10.9), the local

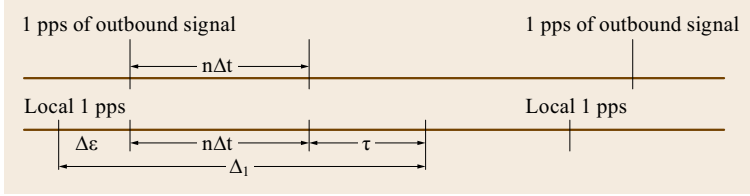


Fig. 10.9 The principle of BDS-1 one-way timing (after [10.9])

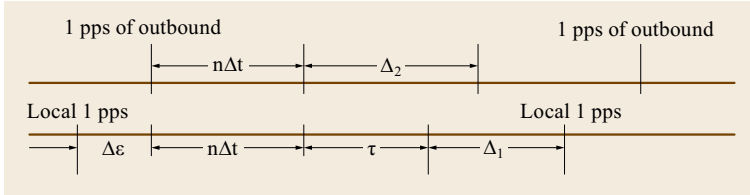


Fig. 10.10 The principle of BDS-1 two-way timing (after [10.9])

clock offset $\Delta\epsilon$ is obtained as

$$\Delta\epsilon = \Delta_1 - (n\Delta t + \tau). \quad (10.5)$$

Here, τ is the signal travel time, which is modeled from the known MCS-satellite and satellite-user distances as well as atmospheric path delays and equipment delay corrections [10.9].

For two-way timing (Fig. 10.10), the user terminal responds to the inbound interrogation signal and the MCS measures the round trip time Δ_2 . From this, the one-way travel time $\tau \approx \Delta_2/2$ is computed in the MCS taking into account various atmospheric and

equipment delay corrections. The resulting value of τ is subsequently sent back to the user. Making use of a local measurement Δ_1 of the time between reception of the interrogation signal and the subsequent 1-pps epoch, the local clock offset is, finally, obtained as

$$\Delta\epsilon = (1 s - \Delta_1) - (n\Delta t + \tau). \quad (10.6)$$

In view of an improved error compensation, the two-way time synchronization achieves an accuracy of 20 ns [10.9], which represents a fivefold improvement over the one-way timing performance.

10.2 BeiDou (Regional) Navigation Satellite System (BDS-2)

In September 2004, the construction of BDS-2 was initiated and a first MEO satellite (then known as COMPASS-M1) was successfully launched in April 2007. It served to protect the frequency filings at the ITU and provided a test bed for the validation of indigenous atomic clocks, precise orbit determination, and time synchronization, as well as other key technologies. Starting with the launch of the first GEO satellite in April 2009, a constellation of 14 operational satellites was deployed in only 3.5 years (Table 10.3).

In December 2012, the BDS-2 entered into official operation and declared the start of its regional service covering latitudes 55°S to 55°N and longitudes 70°E to 150°E [10.10]. The initial service announcement was accompanied by the release of the first open service interface control document (ICD) for single frequency users in the B1 band. An updated version covering also the use of dual-frequency (B1/B2) signals was issued after one year in December 2013 [10.11].

10.2.1 Constellation

The second-generation BeiDou Navigation Satellite System uses a unique constellation design, which combines elements of global systems (such as GPS, GLONASS, and Galileo) with those of purely regional systems (such as the Quasi-Zenith Satellite System (QZSS) and the Indian Regional Navigation Satellite System (IRNSS/NavIC)). The BDS-2 space segment comprises five satellites in geostationary orbit, five spacecraft in IGSO, and four satellites in medium altitude Earth orbit (Table 10.3).

The GEO satellites are positioned at 58.75°E, 80°E, 110.5°E, 140°E, and 160°E, respectively. At least three of them are continuously visible above 10° elevation from any point in the service area, thus, enabling a real-time exchange of information between the control center and the BDS-2 users.

The IGSO satellites operate in circular orbits with an altitude of about 36 000 km and an inclination of 55°.

Table 10.3 Satellites of the regional BeiDou Navigation Satellite System (BDS-2) at the start of operational service. For each satellite, the assigned pseudorandom noise (PRN) code, the international satellite identification number, and the launch date are provided. Satellites in geostationary, inclined geosynchronous and medium altitude Earth orbits are identified by letters “G,” “I,” and “M,” respectively

Satellite	PRN	Int. Sat. Id	Launch	Notes
G1	C01	2010-001A	2010/01/16	140.0°E
G2	–	2009-018A	2009/04/14	Nonoperational
G3	C03	2010-024A	2010/06/02	110.5°E
G4	C04	2010-057A	2010/10/31	160.0°E
G5	C05	2012-008A	2012/02/24	58.75°E
G6	C02	2012-059A	2012/10/25	80.0°E
I1	C06	2010-036A	2010/07/31	~118°E
I2	C07	2010-068A	2010/12/17	~118°E
I3	C08	2011-013A	2011/04/09	~118°E
I4	C09	2011-038A	2011/07/26	~95°E
I5	C10	2011-073A	2011/12/01	~95°E
M1	C30	2007-011A	2007/04/13	Decommissioned
M3	C11	2012-018A	2012/04/29	B3
M4	C12	2012-018B	2012/04/29	B4
M5	C13	2012-050A	2012/09/18	A7
M6	C14	2012-050B	2012/09/18	A8

Like the GEO satellites, they exhibit an orbital period of one sidereal day (i. e., the duration of the Earth rotation relative to the fixed stars, 23 h 56 m), but exhibit a notable inclination with respect to the Earth equator. This results in continuously repeating ground-tracks with a distinct figure-of-eight shape covering a latitude band of $\pm 55^\circ$ (Fig. 10.11). The ground-tracks cross the equator from east to west and the northern part is tra-

versed in a clockwise sense, while the southern loop is traversed in a counter-clockwise direction. Three of the BDS-2 IGSO satellites (I1, I2, and I3) describe a ground track with an equator crossing point at 118°E, while I4 and I5 describe a figure-of-eight centered at 95°E. A new satellite (BEIDOU I1-S/2015-019A) for validation of new BDS-3 signal has been launched in March 2015, filling in the third slot in this ground track.

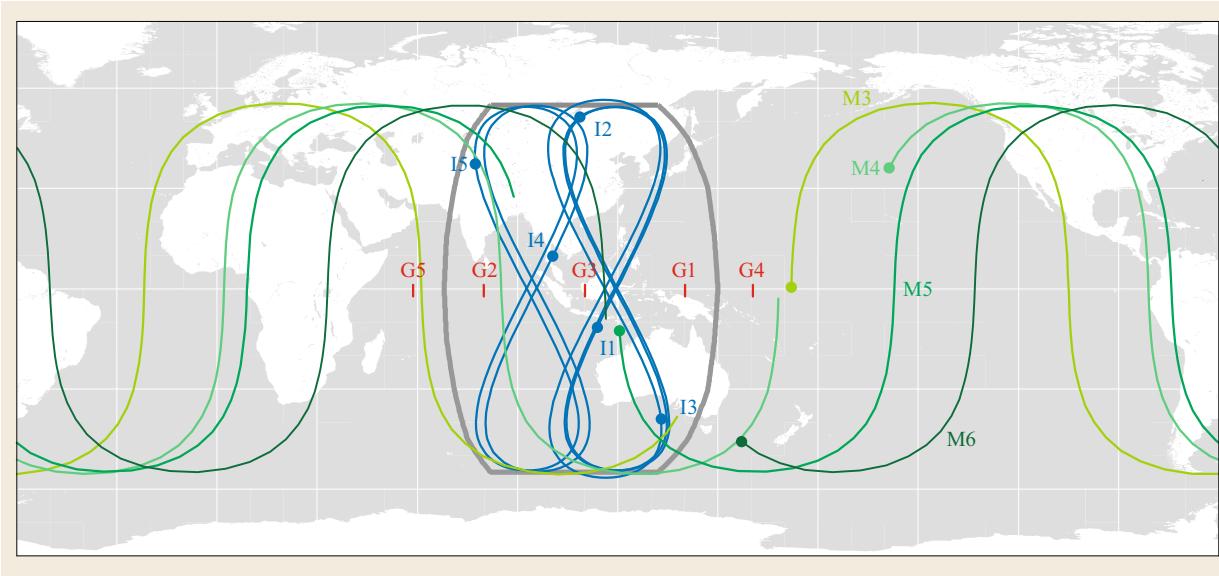


Fig. 10.11 Groundtrack of GEO (red), IGSO (blue) and MEO (green) satellites of the BeiDou Navigation Satellites System (BDS-2) on 1 July 2014. Dots indicate the initial positions at the midnight epoch. The framed region extending roughly from 70°E to 150°E and 55°S to 55°N marks the BDS-2 regional service area as specified in [10.10]

The IGSO satellites in common ground-tracks exhibit a nominal separation of 120° in their arguments of latitude as well as their right ascension of the ascending node (Chap. 3). In this way, they cross the equator at the same geographic longitude with a separation of 8 h.

Even though a combination of GEO and IGSO satellites is commonly considered sufficient and adequate for a purely regional navigation system, the BDS-2 makes complementary use of a small number of MEO satellites. These orbit the Earth at an altitude of 21 530 km (i.e., in between the global positioning system (GPS) and Galileo constellation) and an inclination of 55° . The orbital radius has been chosen such that the satellites complete a total of 13 revolutions within 7 sidereal days, which corresponds to an orbital period of 12h 53m.

The four BDS-2 MEO satellites have been launched in pairs and injected into two different orbital planes, which are separated by 120° in their ascending nodes. In anticipation of a future global extension, the BDS-2 MEO satellite orbits are designed as part of a 24/3/1 Walker constellation [10.12] with 24 satellites evenly distributed in three orbital planes. BeiDou M5 and M6 presently occupy slots 7 and 8 in plane A, while M3 and 4 are placed in slots 3 and 4 of plane B [10.11]. Even though the BDS-2 MEO constellation is still quite limited in size, it increases the average number of visible satellites in the service area and offers additional geometric diversity for improved positioning.

To maintain their nominal position, regular east–west maneuvers with a representative magnitude of about 10 cm/s are conducted by the GEO satellites about once per month [10.13]. North–south corrections, in contrast, are conducted only rarely allowing the buildup of a few degrees of inclination over a period of several years [10.14]. On the IGSO satellites, maneuvers are likewise performed at intervals of about half a year to control the equator crossing longitude [10.15].

10.2.2 Signals and Services

The BDS-2 satellites transmit a total of six signals in three distinct frequency bands: B1 at a center frequency of 1561.098 MHz, B2 at a center frequency of 1207.14 MHz, and B3 at a center frequency of 1268.52 MHz. The B1 and B3 frequencies are offset by about 14 MHz and 10 MHz, respectively, from the E1/L1 and E6 bands of Galileo/GPS, while the B2 center frequency matches that of the Galileo E5b (sub-)band. Each carrier frequency is modulated with two signals in phase quadrature using ranging codes of 2.046 MHz or 10.23 MHz chipping rate (Table 10.4). The resulting signal spectra are illustrated in Fig. 10.12.

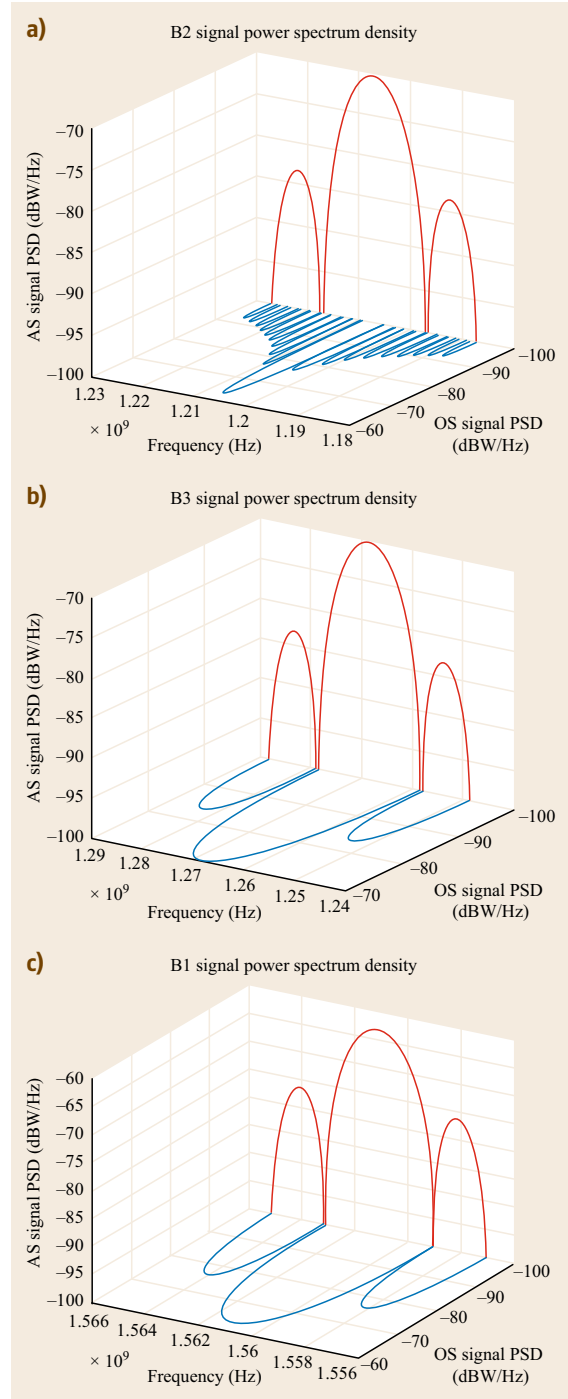


Fig. 10.12a–c Spectral characteristics of BeiDou (BDS-2) signals in the B2 (a), B3 (b), and B1 (c) frequency bands. Note the different scales of the individual plots

The in-phase components of the B1 and B2 signals are assigned to the BDS-2 open service [10.11],

Table 10.4 BDS-2 navigation signals. The BPSK(n) code chip rate is $n \times 1.023$ Mcps

Band	Frequency (MHz)	Signal	Modulation	Service
B1	1561.098	B1-I	BPSK(2)	Open
		B1-Q	BPSK(2)	Authorized
B2	1207.14	B2-I	BPSK(2)	Open
		B2-Q	BPSK(10)	Authorized
B3	1268.52	B3-I	BPSK(10)	Authorized
		B3-Q	BPSK(10)	Authorized

while the remaining four signals are reserved for the authorized service. The open service is specified to offer signal-in-space range errors (SISREs) of better than 2.5 m and a horizontal and vertical positioning accuracy of better than 10 m at a 95% confidence level [10.10] within the BDS-2 service area shown in Fig. 10.11. A minimum received signal power of -163 dBW is ensured for the open service signal on each of the two frequencies.

Apart from the standard positioning and timing service available in the entire service area [10.10], the BDS-2 open service signals also support an improved, satellite-based augmentation systems (SBAS)-like positioning service based on near-real-time corrections transmitted through the GEO satellites. This service is freely available to all users, but limited to a smaller service area centered around the Chinese mainland.

The open service signals employ truncated Gold-codes with a length of 2046 chips and a chipping rate of 2 MHz. Similar to the GPS C/A-code, the BDS-2 open service (OS) codes are generated using a pair of 11-bit shift registers and a full family of codes is obtained through configurable selectors (Fig. 10.13). The native code length of the 11-bit registers amounts to $2^{11} - 1 = 2047$ bits, but the sequence is reset after 2046 chips to obtain a pseudo-random noise (PRN) code of exactly 1 ms duration. Within the OS signal ICD, a total of 37 different open service PRN codes are defined, out of which 1–5 PRNs are reserved for the geostationary satellites. On a given satellite, the same PRN sequence is used for both the B1-I and B2-I OS signals.

Next to the primary ranging code, the open service signal of the IGSO and MEO satellites is multiplied with a secondary, or Neuman–Hofman (NH) [10.16] code. Each chip of the NH-code has a duration of 1 ms and is aligned with the start of the primary code. Depending on the sign of the secondary code chip, either the unmodified primary ranging code sequence or a sign-inverted version is transmitted. As discussed in [10.17, 18], the use of NH-codes helps to decrease narrow-band interference, improves the cross- and autocorrelation properties, and offers an increased robustness in the data bit synchronization. The BDS-2 NH-code (00000 10011 01010 01110) matches that of

the GPS L5-Q signal. It has a length of 20 chips and repeats after 20 ms.

The B1-I and B2-I signals of all BeiDou-2 satellites are, furthermore, modulated with navigation data. For a given satellite, the same data are transmitted on both frequencies, but different formats and data rates are employed for the various types of spacecraft. In case of the MEO/IGSO satellites, the *D1* navigation message is transmitted with a rate of 50 bps (corresponding to a data bit length of 20 ms). The GEO satellites, in contrast, transmit the *D2* message at a 10 times higher rate of 500 bps. Here, the data bit length amounts to just 2 ms, that is, the duration of two consecutive ranging codes.

The overall modulation scheme for the two classes of BDS-2 satellites is illustrated in Fig. 10.14. While improved signal properties at the expense of a lower data rate are favored for the MEO/IGSO satellites, the GEO signal is optimized for the transmission of high data volumes. This different tradeoff reflects the specific role of the GEO satellites, which also provide real-time augmentation data to the BeiDou users as part of the navigation message (Sect. 10.2.3).

The choice of different signal structures for satellites of the same constellation is unique to BDS-2 and requires specific care in the receiver implementation. As pointed out in [10.19] and [10.20], inconsistent interpretations of the NH-code sign may result in a half-cycle intersatellite-type biases (ISTB) when forming double-difference carrier-phase observations between GEO and non-GEO BDS-2 satellites using different receiver types. Further aspects of BeiDou receiver design and the impact of the NH-code on acquisition and tracking of the MEO/IGSO signals are discussed in [10.21] and [10.22].

Even though only the B1-I and B2-I signals are officially declared as open service signals, basic properties of the (authorized) B3-I signal have also become public through inspection with high-gain antennas [10.23, 24]. The signal employs a primary ranging code of 10 230 chips with a chipping rate of 10.23 MHz and a resulting length of 1 ms as well as a 20-bit secondary code. Knowledge of the ranging code, which can be generated by linear feed-back shift registers, has enabled various manufacturers to provide geodetic-grade

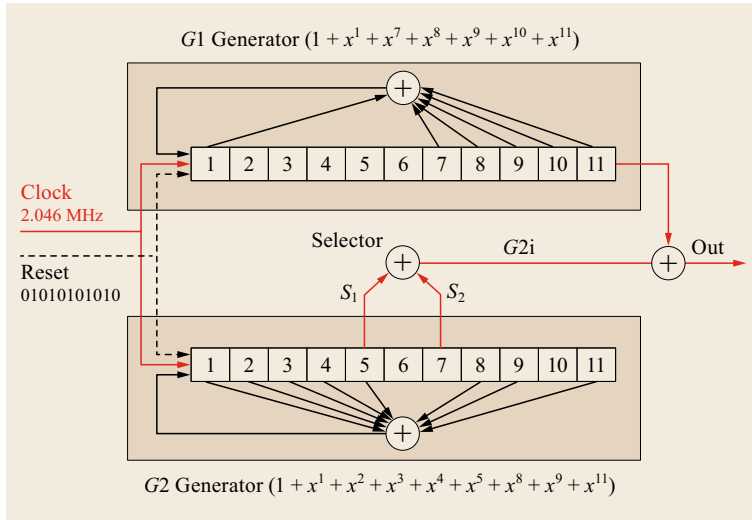


Fig. 10.13 Code generator for the B1-I and B2-I open service signals (after [10.11])

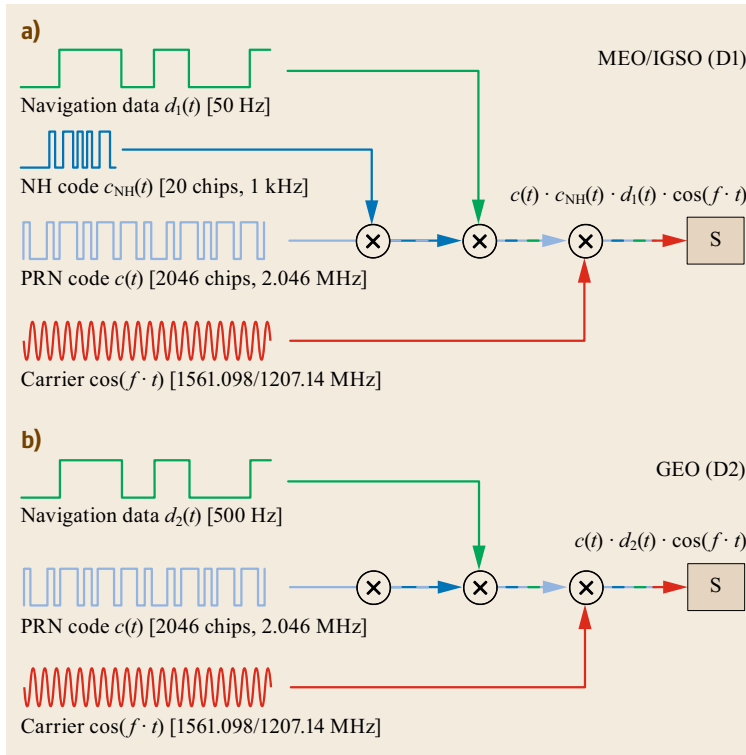


Fig. 10.14a,b Signal structure of the BeiDou-2 open service signals for MEO/IGSO (a) and GEO satellites (b)

triple-frequency BeiDou receivers. Access to more than two frequencies in BeiDou paves the way for various advanced GNSS processing techniques [10.25–28] with applications in surveying, precise point positioning and geodesy. It must be kept in mind, though, that tracking of the B3-I signal is not officially endorsed and may be inhibited at any time using explicit encryption of the B3-I ranging code.

10.2.3 Navigation Message

The D1 navigation message broadcast by the MEO and IGSO satellites resembles the legacy navigation message of GPS transmitted with the L1 C/A code signal. It employs the same data rate and provides basic navigation information such as almanac and ephemeris data for acquisition of visible satellites as well as positioning

and timing. On top of these data, the high-rate D2 message of the BeiDou GEO satellites contains additional augmentation service information such as BDS integrity, differential corrections, and ionospheric grid data.

D1 Message Structure

The D1 navigation message is structured into a superframe made up of individual frames and subframes (Fig. 10.15). Each subframe is composed of 300 bits and takes 6 s to transmit. Five subframes make up a full frame, which is of 30 s duration. The superframe, finally, comprises 24 frames (i.e., one less than GPS) with a total of 36 000 bits transmitted in 12 min. Subframes 1–3 contain fundamental navigation data such as orbit and clock parameters for the transmitting satellite as well as ionospheric model coefficients. This information is updated at the start of each hour and repeated once every 30 s within this period.

Subframes 4 and 5 are commutated, that is, different pages are transmitted within consecutive frames. They contain the almanac information for up to 30 satellites (in pages 1–24 of subframe 4 and pages 1–6 of subframe 5) as well as constellation health data, the BDS time offset from UTC as well as the time offsets between BDT and other GNSS time scales (pages 7–10 of subframe 5). The entire information is repeated once every 12 min.

The 300 bits of each subframe are composed of 10 elementary 30-bit words. Each halfword contains 11 data bits as well as 4 parity bits encoded with a Bose–Chaudhuri–Hocquenghem (BCH) [10.29] code enabling single-bit error correction. Furthermore, the data and parity bits from two halfwords with each word are interleaved to protect against possible burst errors during the transmission.

D2 Message Structure

The D2 message is transmitted at a 10 times higher data (500 bps) than the D1 message and uses a different scheme to transmit the low-rate basic navigation data along with high-rate augmentation data. The superframe is composed of 120 frames, which are divided into five subframes of 300 bits length. A single subframe has a length of 0.6 s and a full frame is broadcast in 3 s. The entire superframe contains a total of 180 000 bits and takes 6 min to transmit (Fig. 10.16).

Subframe 1 provides basis navigation data such as orbit and clock information of the transmitting satellite and ionospheric model parameters in a series of 10 commutated pages. Apart from the different layout, the data are identical to subframes 1 and 3 of the D1 message. As in the case of the non-GEO satellites it, thus, takes a total of 30 s to receive the full ephemeris data of a GEO satellite despite the higher transmission rate. Within a superframe, the same set of 10 pages is repeated 12 times. The ephemeris information is complemented by almanac data (for up to 30 satellites) in pages 37–60 and 95–100 of subframe 5. One full almanac is transmitted per superframe and the information is repeated every 6 min.

Information on the integrity of BDS-2 satellites as well as differential correction data are provided to the users in subframes 2 and 3 of the D2 navigation message. Other than in traditional SBAS systems (Chap. 12), equivalent clock corrections are provided in the BDS-2 augmentation data. The scalar correction values Δt account for the combined effects of broadcast orbit and clock offset errors and are intended for users in China and adjacent areas. The user adds the value of Δt to the observed pseudorange. Updated values are provided once every 18 s. Within this period, six

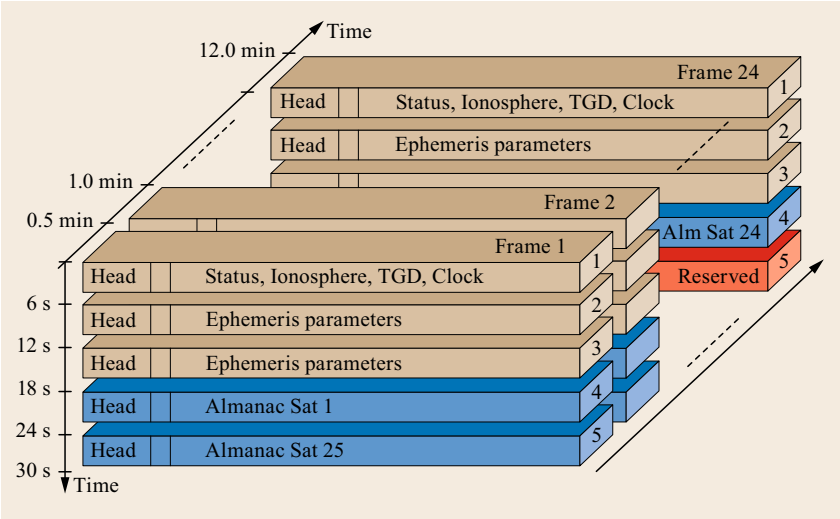


Fig. 10.15 Structure of the BeiDou D1 navigation message transmitted by the MEO and IGSO satellites. Arrows indicate the transmission order of individual subframes

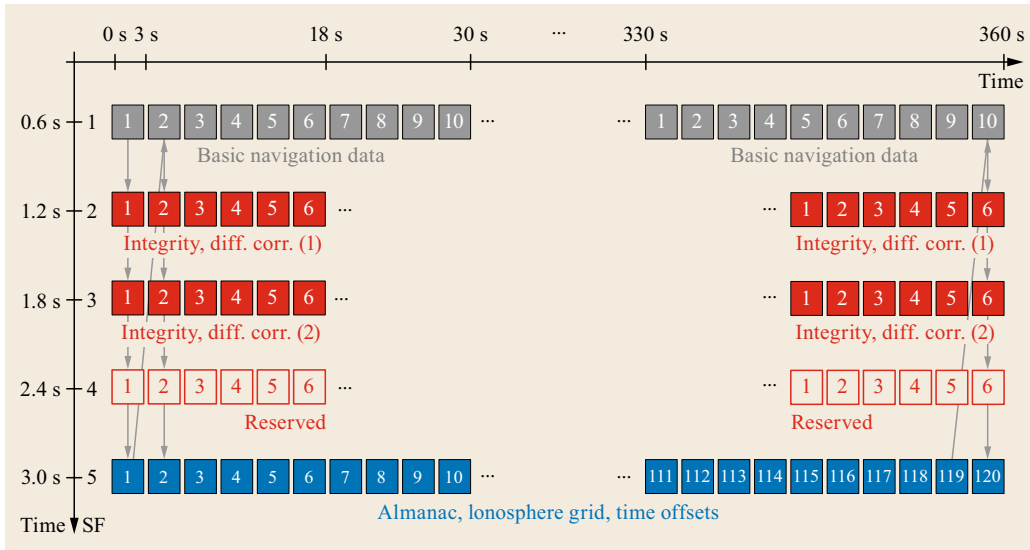


Fig. 10.16 Structure of the BeiDou D2 navigation message transmitted by the GEO satellites (after [10.30])

commutated pages of subframes 2 and 3 are transmitted, which provide integrity information and equivalent clock corrections for up to 18 different satellites. This covers the expected maximum number of BeiDou satellites jointly visible in the service area for augmentation data.

The differential ephemeris corrections are complemented by low-latency ionospheric corrections for single-frequency users provided in subframe 5. These data cover a service area from 70°E to 145°E longitude and 7.5°E to 55°E latitude and are provided in two subgrids of 5° × 5° resolution which are shifted by 2.5° in latitude (Fig. 10.17). In accord with the length of a superframe, the ionospheric grid data are updated

once every 360 s. Depending on the availability and distribution of monitoring stations, the provision of valid correction data may be confined to a subset of points in the overall grid (typically the Chinese mainland).

Ephemeris Parameters and Models

The broadcast orbit and clock parameters as well as the standard ionospheric correction parameters in BDS-2 have been defined in close correspondence with those of the legacy GPS navigation message. A total of 16 orbit parameters for a perturbed Keplerian model and three parameters of a quadratic clock model are used to describe the satellite motion and clock offset variation within the ephemeris validity interval. Ionospheric path

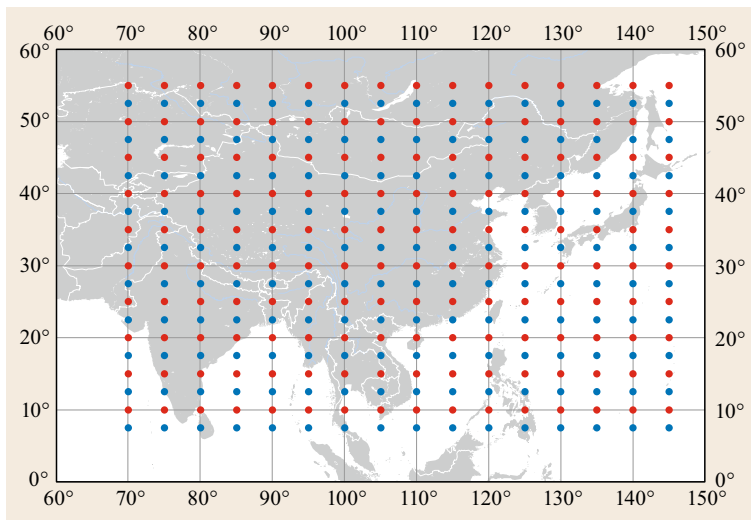


Fig. 10.17 Ionospheric correction data in the BeiDou-2 D2 navigation message are provided for two interleaved grids of 5° width with a 2.5° offset. The northern grid (blue) is transmitted in the first 3 min, followed by the southern grid in the second half of a superframe

delays are described through a Klobuchar-type model with eight parameters.

Despite obvious similarities, the models employed with the BDS-2 broadcast navigation message differ in various aspects from those of GPS [10.11, 30]. All information is referred to BDT (Sect. 10.2.8) and the BeiDou coordinate system (BDC) connected to China Geodetic Coordinate System (CGCS) 2000 (Sect. 10.2.7). Along with that, slightly different values for the gravitational coefficient (GM_{\oplus}) and the Earth rotation rate (ω_{\oplus}) are employed in BeiDou.

With the above exceptions, the orbit model of BDS-2 matches that of GPS for MEO and IGSO satellites, but a different interpretation of the orbit parameters applies for the GEO satellites (Sect. 3.3.2). Since the geostationary orbits exhibit a very low (and potentially zero) inclination, large residuals and divergence problems may be encountered when determining the broadcast elements. To cope with this situation, a reference plane with a 5° inclination angle to the equator is adopted in the orbit determination of the GEO satellites. A complementary 5° rotation must, therefore, be applied when calculating the position and velocity for these satellites relative to the Earth equator.

Satellite clock offsets from BDT are described through a second order polynomial

$$\Delta T = a_0 + a_1(t - t_{oc}) + a_2(t - t_{oc})^2 + \Delta t_{rel} \quad (10.7)$$

with coefficients a_0 , a_1 , and a_2 providing offset, drift, and drift rate at the reference epoch t_{oc} , as well as periodic relativistic corrections

$$\Delta t_{rel} = -\frac{2}{c^2}(\mathbf{r}^\top \mathbf{v}) \quad (10.8)$$

depending on the spacecraft position \mathbf{r} and velocity \mathbf{v} in much the same way as for GPS and Galileo. However, the clock offsets are referred to single-frequency B3 observations rather than an ionosphere-free dual-frequency combination. All users of the open service BDS-2 signals must, therefore, account for additional satellite group delay corrections. For processing of single-frequency B1-I observations, a timing group delay parameter TGD_1 provided in the navigation message has to be added to the B3 clock offset value, while a linear combination of TGD_1 and TGD_2 is required for dual-frequency (B1-I/B2-I) observations [10.31].

For ionospheric correction of single-frequency observation users, the BDS-2 broadcast navigation message makes use of a Klobuchar-style thin layer correction model (Sect. 6.3.4). It involves a total of eight coefficients ($\alpha_0, \dots, \alpha_3$ and β_0, \dots, β_3) describing the amplitude and period of the daytime vertical total electron content variation. The ionospheric model is constructed in geographic latitude and longitude, rather

than geomagnetic coordinates as used in GPS. The coefficients of this model are transmitted by all BDS-2 satellites and updated at hourly intervals. Even though the model is conceptually valid on a global scale, best results are obtained across the Chinese mainland due to currently limited distribution of monitoring stations [10.32].

10.2.4 Space Segment

The BeiDou regional navigation satellite system is composed of three groups of spacecraft, namely the GEO, IGSO, and MEO satellites. The satellites have a specified lifetime of eight years. The platform and navigation payload of all three satellites types are essentially similar, but various complementary payloads are accommodated on the GEO spacecraft.

Satellite Platform

The BDS-2 MEO/IGSO satellites (Fig. 10.18) adopt the DongFangHong-3 (DFH-3; *The East is Red*) platform, and a slightly modified version (DFH-3A) is employed for the GEO satellites spacecraft [10.14, 33]. The three-axis stabilized DFH-3 bus has been developed by the China Academy of Space Technology (CAST), Beijing, in cooperation with Messerschmitt-Bölkow-Blohm (MBB), Germany, and was used for various geostationary communications satellites since 1994 [10.34]. It is equipped with an apogee boost motor and liquid propulsion system for initial orbit insertion and orbit keeping purposes.

At a size of about $2.4 \times 1.7 \times 1.7 \text{ m}^3$ (GEO) and $2.0 \times 1.7 \times 1.7 \text{ m}^3$ (MEO/IGSO), the spacecraft have different masses of around 2.5 t at lift-off. The platform comprises subsystems for thermal control, power supply and distribution, tracking, telemetry and commanding (TT&C), as well as attitude and orbit control. The solar arrays with a total area of more than 20 m^2 are based on silicon cells (MEO/IGSO satellites) and GaAs/Ge cells (GEOs) providing a minimum power of 2 kW and 2.5 kW, respectively. For power storage, all BDS-2 satellites are equipped with NiMH batteries offering capacities of 40–60 Ah [10.14].

Attitude control is performed through a combination of Earth and Sun sensors as well as four reaction wheels. These wheels are unloaded during station-keeping maneuvers or using magnetorquers to avoid unnecessary orbital perturbations. Outside the eclipse season, a continuous yaw-steering (Sect. 3.4) is performed by the MEO and IGSO satellites to maintain the antennas pointing to the center of the Earth while orienting the solar panels to the Sun [10.35]. The Earth-pointing accuracy is typically better than 0.25° and the nominal solar panel orientation is maintained with an

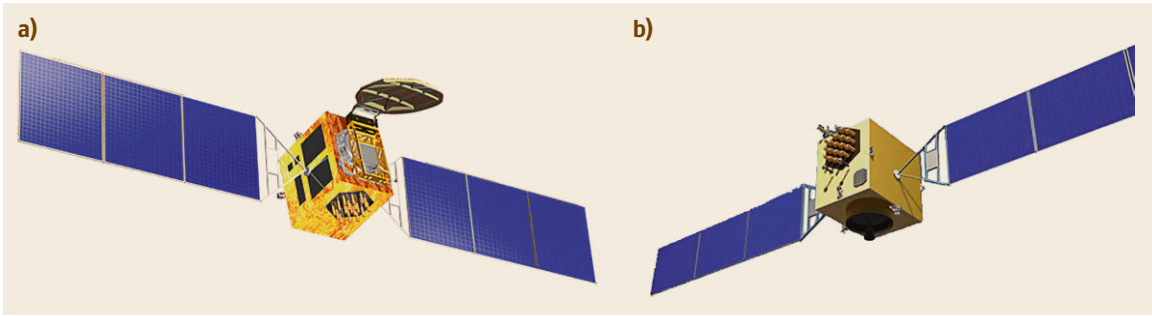


Fig. 10.18a,b GEO (a) and MEO/IGSO (b) satellites of the regional BeiDou Navigation Satellite System (BDS-2). Reproduced with permission of Beijing Satellite Navigation Center

uncertainty of less than 5° [10.14]. To avoid the need for rapid yaw-slews near Sun-spacecraft-Earth collinearity, an orbit-normal attitude is employed, whenever the elevation of the Sun above the orbital plane is less than about 4° [10.36, 37]. The orbit-normal mode is also adopted for all BeiDou GEO satellites. Measured yaw angles provided in the onboard telemetry were found to match the nominal attitude at a level better than $0.5\text{--}1^\circ$ [10.38].

Satellite Payload

The RNSS navigation payload of the BeiDou satellites is mainly composed of the time and frequency subsystem as well as the navigation processor and signal generation unit.

The time and frequency subsystem is used to generate, maintain, and calibrate the primary reference frequency. It consists of Rubidium atomic frequency standards (RAFSs; Fig. 10.19) as well as supporting equipment such as a frequency multiplier, reference frequency synthesizer, and power-division/amplification network. The four clocks on each spacecraft comprise one active clock, a hot backup, and two cold backups. While the majority of clocks employed in the BDS-2 constellation is of indigenous origin, a limited amount of European RAFS has been procured for the BeiDou program and is used alongside the Chinese clocks on the various spacecraft [10.8, 14, 39]. The RAFS performance is further described in Sect. 10.3.2.

Following the modulation of the carrier with the ranging code and navigation message, the navigation signals are filtered and amplified in a traveling wave tube amplifier (TWTA). The phased-arrays antenna used for transmission of the signals covers a total of three frequency bands (B1, B2, and B3). Similar to GPS, a slightly higher gain in the off-boresight direction is used to compensate the increased distance and free-space loss for users at low elevations. The desired shape of the gain pattern is achieved by phase coherent combination of a central part with 6+1 helix antenna el-

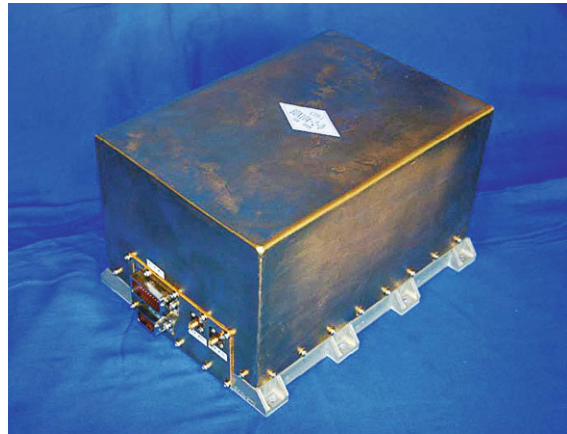


Fig. 10.19 Rubidium atomic frequency standard for BeiDou-2 satellites (after [10.40]). Reproduced with permission of Beijing Satellite Navigation Center

ements as well as an outer ring of 12 individual helix antennas.

The upload of navigation information from the ground is handled by the uploading and ranging segment, which comprises an L-band uploading receiver and a spread spectrum ranging receiver for calibration of the two-way time delay.

In addition to the RNSS payload, which is common to all BeiDou-2 spacecraft, the geostationary satellites are equipped with an RDSS payload comprising an L-band/C-band inbound transponder and C-band/S-band outbound transponders for BDS-1-type navigation and short messaging (Sect. 10.1). Furthermore, the GEO satellites host a C-band transponder for time synchronization and data transmission between the ground control center and monitoring stations.

All BDS-2 satellites are equipped with a laser retroreflector array (LRA), which is used for precise orbit determination, and to assist the time comparison between satellites and ground stations. The LRAs were developed by Shanghai Observatory and consist

of a planar array of individual prisms with a diameter of 33 mm. To account for the large distance, an LRA with a size of about $50 \times 40 \text{ cm}^2$ and a total of 90 prisms is used onboard of GEO and IGSO satellites, whereas a smaller, 42-prism LRA is sufficient for the MEO spacecraft. Furthermore, the LRAs are slightly tilted on the GEO satellites, to improve the return signal strength for satellite laser ranging stations in China [10.41].

Complementary to satellite laser ranging (SLR), various BeiDou MEO/IGSO satellites support laser time transfer (LTT) through a dedicated onboard detector and timer connected to the atomic clocks [10.42]. The LTT system developed by the Technical University of Prague and Shanghai Astronomical Observatory [10.43] is based on a single photon avalanche diode (SPAD) and enables measurements of the ground-to-satellite signal travel time with subnanosecond (i.e., a few centimeter-level) precision. Independent of GNSS observations or traditional two-way time transfer (TWTT) equipment, LTT enables the precise monitoring and synchronization of the atomic frequency standards of the BeiDou satellites.

10.2.5 Operational Control System

The Operational Control System (OCS) of BDS, like that of other GNSSs, is a key segment for BDS operation. The research on the OCS of BDS demonstration system started as early as the 1980s. The OCS of BDS-2 has been in place since 2007, when the first BDS-2 satellite was launched.

The OCS of BDS provides command, control and operation capabilities for the three kinds of satellite constellations. The main functions of BDS OCS are:

- To establish and maintain coordinate reference datum
- To maintain time reference datum
- To measure time synchronization between satellites and ground stations
- To manage precise orbit determination and prediction
- To predict satellite clock offset
- To deal with augmentation data
- To process RDSS information
- To monitor, process, and predict ionospheric delay
- To monitor integrity
- To upload and download navigation message.

The workflow of the OCS can be simply described as collecting data related to the satellites and the ground stations, processing and analyzing the collected data, managing communication between satellites and ground stations, uploading operational commands to

satellites, and broadcasting navigation messages to the users.

In the course of the BDS development, the OCS has evolved along with the increased number of satellites and complexity of the constellation. While the OCS in BDS-1 only controlled the two GEO satellites and one backup, the OCS in BDS-2 has to control three different types of satellites (GEO, IGSO, and MEO). The information, which the OCS has to process, has been enlarged dramatically. In addition, the services have been extended from RDSS to RDSS plus RNSS. Finally, the calibration stations have been replaced by the modern monitoring stations. There are high database and calibration stations with barometric altimeter for positioning in the OCS of BDS-1. The OCS of the BDS-2, however, no longer needs high database and some of the new monitoring stations are equipped with laser ranging system.

At present, the OCS of BDS-2 is composed of 1 MCS, 7 monitoring stations of Type A, which are mainly used for orbit and ionospheric delay monitoring, 22 monitoring stations of Type B, which are mainly used for augmentation service and integrity service, and 2 time synchronization/uploading stations (Fig. 10.20).

The frequency bands of the BDS-2 RDSS feeder and service links are the same as in BDS-1. The L-band is used in the BDS-2 RNSS service such as navigation message upload and download (Fig. 10.21).

The OCS performance in BDS-2 is reflected by the general accuracy of the satellite orbits, the time datum, and the coordinate datum. The orbit determination accuracy as obtained from the regional monitor network is about 0.2 m, 1.2 m, and 0.6 m in radial, along-track, and orbital normal (N) directions, respectively. Satellite laser range measurements agree with the determined orbits to better than 1 m root-mean-square (rms) [10.38]. The user ranging error (URE) of 2-h orbit prediction is 1 m for GEO satellite, while UREs of IGSO and MEO orbit prediction are about 1 m at 6 h and 5 m at 16 h, respectively. The accuracy of satellite clock prediction ranges from 1.4 ns at 2 h to 12 ns at 10 h. Finally, the accuracy of ionospheric delay corrections by the broadcast Klobuchar model is better than 75%.

The updating strategy of the OCS involves mainly six parts: the least-squares adjustment of a polynomial model for the satellite clock using robust estimation [10.8]; generation of the ephemeris, which is presently updated each hour; adjustment of the coefficients for the Klobuchar ionospheric model, which is based on data from 32 monitoring stations; performance of the RDSS two-way service with a response time of less than 1 s; transfer of RDSS messages of up to 120 Chinese characters of 14 bit each; and issue of integrity warnings with a response time of less than 8 s.

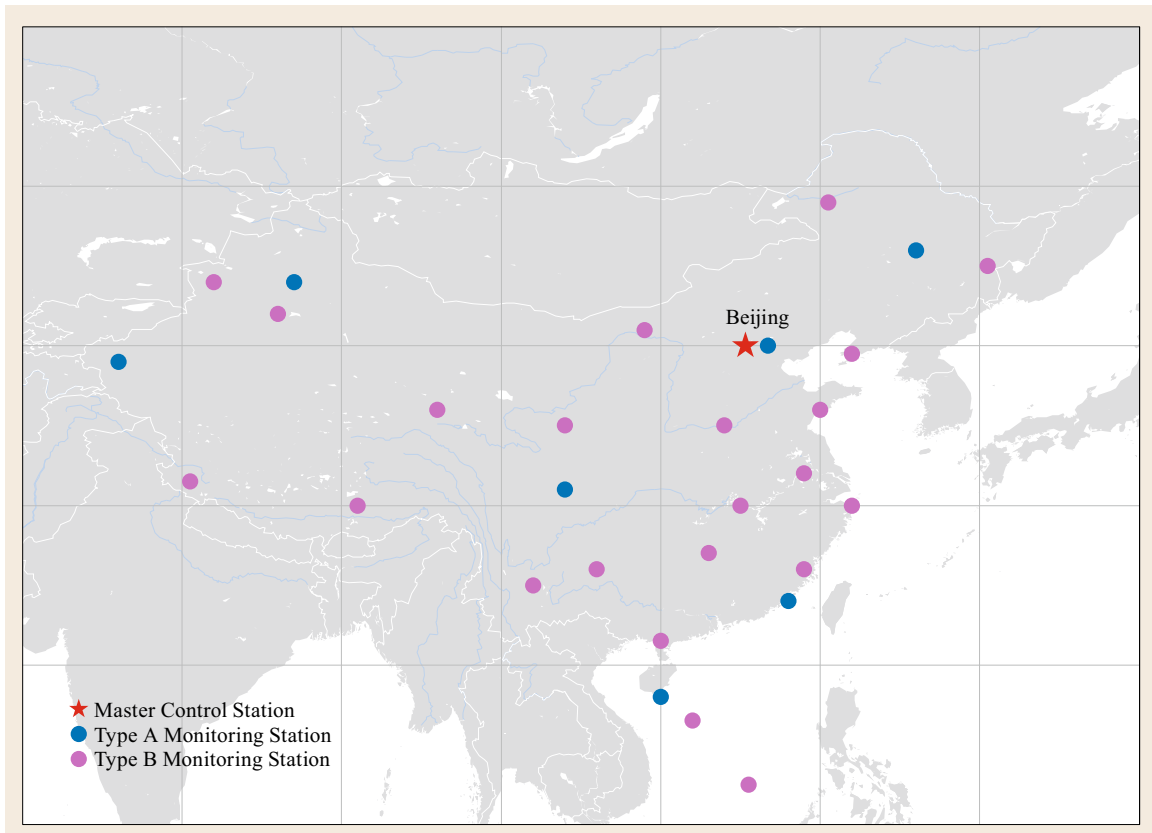


Fig. 10.20 Location of BeiDou Master Control Station and the type-A/B monitoring stations. Information courtesy of Beijing Satellite Navigation Center

As part of the ongoing BeiDou evolution to a fully global service, the OCS will be adapted to operate the next-generation of BDS satellites (MEO/IGSO/GEO) in an optimal and robust manner. The ground segment will be extended to upload navigation messages for the entire satellite constellation, to maintain an interoperable time datum based on reliable, stable, and accurate clocks, and to maintain the BeiDou coordinate system in real time using multi-GNSS observations.

10.2.6 BeiDou Satellite-Based Augmentation System

Unlike other SBASs (Chap. 12), which are separated from the navigation satellite systems, BDS and its OCS include the BeiDou Satellite-based Augmentation Services (BDSBAS, [10.44]) as an integral part.

After the BDS-1 was established, the first-generation of BDSBAS (BDSBAS-1) was embedded in 2003. It broadcasted augmentations for the GPS navigation signal using the GEO satellites. More than 20 monitoring stations were established for tracking

and integrity monitoring. In each monitoring station, three receivers and one atomic clock were assembled. All the collected monitoring data were transmitted to the MCS, where the precise satellite orbit, grid ionospheric corrections, and the equivalent satellite clock corrections were processed. The corresponding corrections, the user differential range error (UDRE), and grid ionospheric vertical error (GIVE) were generated.

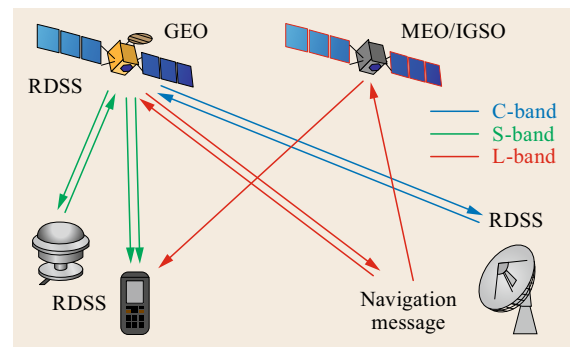


Fig. 10.21 Signal flow of satellites and monitoring stations

It was shown that the precision of differential positioning based on GPS L1 C/A was about 3 m, and the integrity, continuity, and availability were also improved.

In BDS-2, the basic navigation service and augmentation service have also been integrated [10.11], for which technical tests were started in December 2012. The differences of the BDSBAS-1 and BDSBAS-2 are as follows:

- The augmentation service of only GPS L1 C/A in BDSBAS-1 has been extended to include both GPS L1 C/A and BDS BII, and possibly Galileo and other GNSSs in the future
- The service coverage area is enlarged in BDSBAS-2, because all five GEO satellites broadcast the augmentation information, as opposed to only two GEO satellites in the BDSBAS-1
- In the BDSBAS-1, the integrity was not included in the augmentation information, but it is included in the BDSBAS-2
- The number of monitoring stations in BDSBAS-2 has been increased to more than 30
- Scalar *equivalent satellite clock errors* were derived and broadcast in BDSBAS-1, while satellite clock errors and satellite orbit errors are handled separately in the BDSBAS-2.

The workflow of calculating the equivalent satellite clock errors is as follows:

- Geometric distances between the satellite and monitoring stations are calculated based on the broadcast ephemeris and the satellite clock corrections as well as the known coordinates of the monitoring stations.
- The differences between the geometric distances and measured pseudoranges which are corrected by the ionospheric influences are calculated; thus, the pseudorange residuals are obtained.
- The equivalent satellite clock correction is obtained as the average of the residuals corresponding to this satellite.

The augmentation data of BII signal are monitored by the monitoring stations and then the monitored data are transmitted to the MCS and processed there. The corrections of ephemeris, satellite clocks (SCs), and grid ionosphere as well as UDRE/GIVE and integrity information are generated in the MCS and modulated in the navigation message. The augmentation information together with the normal navigation message is uploaded to the GEO satellites through the communication link between the MCS and the GEO satellites, on

which the augmentation messages are modulated into the downlink signals and broadcast to the users. In the new test, the precision of differential positioning by BII of BDS is better than 3 m.

10.2.7 Coordinate Reference System

The BDC system is connected to the China Geodetic Coordinate System 2000 (CGCS2000), which itself is aligned to the International Terrestrial Reference System (ITRS) [10.45].

CGCS2000 is realized by the China Terrestrial Reference Frame (CTRF). The definition of this coordinate system follows the criteria outlined in the 1996 conventions [10.46] of the International Earth Rotation and Reference Systems Service (IERS). The next generation of CGCS may be renewed following the new IERS conventions issued in 2010 [10.47].

CGCS2000 is a right-handed, orthogonal system. Its origin is the center of the mass of the Earth including the oceans and the atmosphere. Its scale is that of the local Earth frame, in the meaning of a relativistic theory of gravitation [10.45, 46]. The orientation is initially given by the orientation of the Bureau International de l'Heure (BIH) Terrestrial System at 1984.0, and the time evolution of the orientation is ensured by using a no-net-rotation condition with respect to the horizontal tectonic motions over the whole Earth. The unit of the length is meter. Its z -axis is the direction of the IERS Reference Pole (IRP). This direction corresponds to the direction of the BIH conventional terrestrial pole (CTP) at epoch 1984.0. The x -axis is the intersection of the IERS reference meridian (IRM) and the plane passing through the origin and normal to the z -axis. The IRM is coincident with the BIH Zero Meridian at epoch 1984.0. The y -axis, finally, completes a right-handed, Earth-centered Earth-fixed (ECEF) orthogonal coordinate system.

The CGCS2000 origin also serves as the geometric center of the CGCS2000 ellipsoid and the z -axis serves as the rotational axis of this ellipsoid of revolution. Parameters of the CGCS2000 reference ellipsoid are listed in Table 10.5.

Table 10.5 Fundamental parameters of the CGCS2000 system

Parameter	Value
Semimajor axis	$a = 6\,378\,137.0\text{ m}$
Flattening	$f = 1/298.257222101$
Gravitational coeff. (incl. atmosphere)	$GM_{\oplus} = 398\,600.4418 \cdot 10^9\text{ m}^3/\text{s}^2$
Angular velocity	$\omega_{\oplus} = 7.292115 \cdot 10^{-5}\text{ rad/s}$

10.2.8 Time System

The time reference for the BeiDou (Regional) Navigation Satellite System is BeiDou Time, BDT [10.8]. BDT is a continuous navigation time scale without leap seconds and with the SI second as its basic unit. BDT is commonly represented by the BeiDou week number (WN) and the seconds of week (SoW), ranging from 0 to 604 799. The zero point of BDT (i. e., WN = 0, SoW = 0) is January 1, 2006 (Sunday) UTC 00h00m00s. Similar to GPS and Galileo Time, BDT is aligned to UTC except for an integer second offset caused by the accumulated leapseconds.

The basic tasks of time keeping and timing service of the BeiDou system are to provide time and frequency signals in real time with continuity, stability, high accuracy, and reliability. Here, *continuity* reflects that time (frequency) is differentiable and the ability of the time system to run without interruption; *stability* refers to the variation of frequency over time and is commonly expressed by the Allan deviation (ADEV); *accuracy* means the consistency of the time (frequency) signal with the nominal value, which is often denoted by relative time bias (frequency bias); and *reliability* means the ability to provide the time and frequency signals under given conditions over the envisaged period of operation. In fact, there is no single atomic clock, which can provide time and frequency signals meeting the above criteria. Multiple Hydrogen maser clocks are, therefore, employed to collectively provide the BeiDou system time and frequency with the desired performance and reliability.

BDT is maintained by the Time and Frequency System (TFS) and aligned to the UTC realization UTC(BSNC) of the Beijing Satellite Navigation Center (BSNC). The TFS is mainly composed of five parts shown in Fig. 10.22. The clock ensemble (CE) comprises about 10 Hydrogen masers, which are used to

form a *composite clock* in a robust estimation process. The intermeasurement element (IME) provides measurements of the original time and frequency signals from the clock ensemble and outputs the clock differences both in time and frequency in a regular pattern. The outer comparison element (OCE) provides the deviation of BDT from other time scales, especially from UTC and that of NTSC (National Timing Service Center of Chinese Academy of Science). The data processing element (DPE) carries out the calculation with the given algorithm based on all the information from IME and OCE, to give a relative uniform time scale, which is called BDT and works as the time reference for the whole navigation system. The signal generation element (SGE), finally, exerts frequency adjustments to the signals of the master clock (MC), and generates physical time and frequency signals for the MCS.

The algorithm performed in BDT calculation is well designed to form a good composite clock. The frequency offset, drift, and instability of each clock have been taken into account. The weights of clocks are determined by their Allan variances in which the frequency drift is taken off. The equivalent weights are also used based on the robust estimation principle [10.48, 49]. In order to be as consistent as possible with UTC, BDT may be steered with an interposed frequency adjustment after a period of time (more than 30 days) if the bias between the BDT and UTC is beyond the appointed threshold, but the quantity of the interposed frequency adjustment is not allowed to be more than $5 \cdot 10^{-15}$.

In order to align BDT with UTC, a time and frequency transfer chain between the MCS and the NTSC has been established. It makes use of two-way satellite time and frequency transfer (TWSTFT) via the geostationary BeiDou satellites as well as BDS and GPS common view (CV) observations [10.50]. Fiber chains between BSNC and National Institute of Metrology

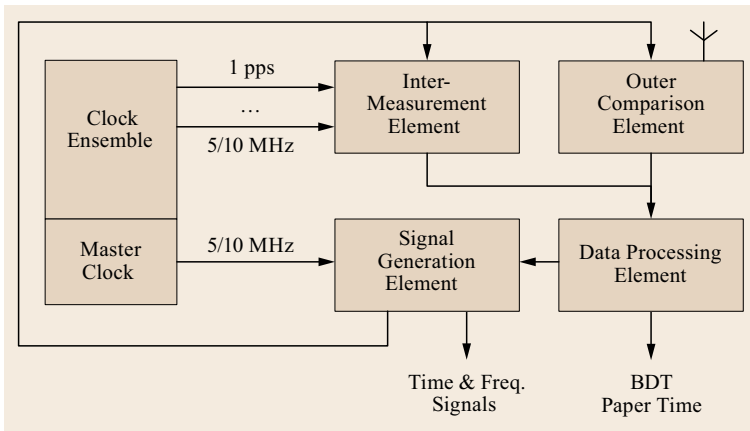


Fig. 10.22 The composition of the time and frequency systems (after [10.8])

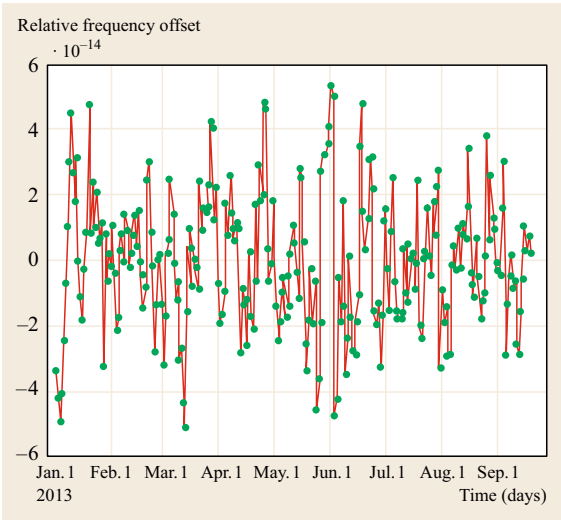


Fig. 10.23 Frequency bias of BDT relative to UTC(BSNC)

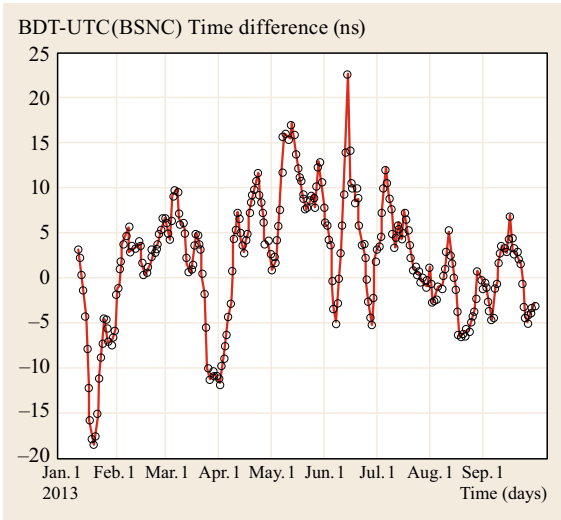


Fig. 10.25 Difference of BeiDou system time with respect to UTC(BSNC)

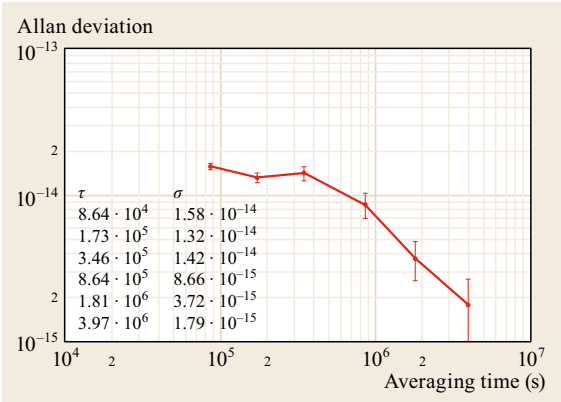


Fig. 10.24 Frequency stability of BDT shown by ADEV

Table 10.6 BeiDou system time performance

Patameter	Value
Time (frequency) accuracy	$< 1.0 \cdot 10^{-13}$
Time (frequency) stability	$< 2.0 \cdot 10^{-14} / 1d$
	$< 1.0 \cdot 10^{-14} / 7d$
Time deviation BDT-UTC	$< 100 \text{ ns (modulo 1 s)}$

(NIM) as well as BSNC and NTSC will be established. Apart from the indirect links, a direct determination of the BDT-UTC time offset through TWTFVT and GNSS CV measurements between the BSNC and the Bureau International des Poids et Mesures (BIPM) is under preparation. For further reference, Figs. 10.23 and 10.24 show the frequency offset and ADEV of BDT in comparison with the realization of UTC pro-

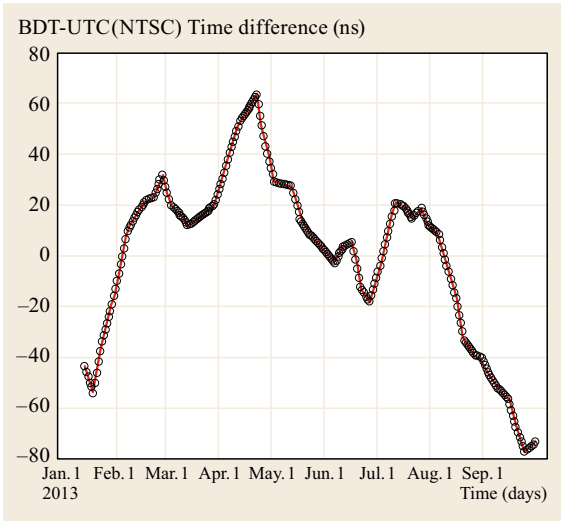


Fig. 10.26 Difference of BeiDou system time with respect to UTC(NTSC)

vided by the BSNC. The overall offset between BDT and UTC(BSNC) as well as UTC(NTSC) (which is closely aligned with UTC and GPS time) is shown in Figs. 10.25 and 10.26 over a 10-month period in 2013. All graphs are based on data of the BeiDou MCS.

The current performance of BDT is summarized in Table 10.6. It may be noted that the performance of BDT after running several years is not as good as its original state shown in [10.8].

10.3 Performance of BDS-2

During the construction of BDS-2, the performance of satellite clock, ranging, positioning accuracy, and reliability has been validated gradually. Tests performed by different communities [10.38, 51] demonstrate that the performance of BDS-2 in positioning, navigation, and timing (PNT) has reached or even exceeded the design specification.

10.3.1 Service Region

Broadcast ephemeris date for January 22–29, 2013 was used in [10.51] to assess the number of visible satellite and position dilution of precision (PDOP) [10.52, 53] for global users and the Asia-Pacific area shortly after the start of the regional operational service.

For areas between latitude 70°S–70°N and longitude 40°E–180°E, the number of visible satellite is more than five (Fig. 10.27) and the PDOP is smaller than 12 (Fig. 10.28), which satisfies the basic navigation requirements. More reliable navigation is available between latitude 60°S–60°N and longitude 70°E–150°E, where the number of visible satellite is more than seven and the PDOP is less than five. For areas between latitude 50°S–50°N and longitude 85°E–135°E, the number of visible satellite is finally larger than eight and the PDOP is between 2 and 3. Higher accuracy and a more reliable navigation service can be provided in these areas.

The number of visible satellites is above seven in China, and the availability (as defined by a PDOP value of less than <6) exceeds 97.5%. The service region of BeiDou system reaches the design requirement, and the availability of BeiDou system is close to 100% in the designated coverage area (Fig. 10.29).

10.3.2 Performance of Satellite Clocks

The performance of BeiDou SCs directly affects the accuracy and reliability of the users' positioning and navigation results. From the measurements between the BeiDou satellites and the ground uplink stations, the on-board clock offsets from BDT and the clock stabilities can be evaluated. In addition, the performance can also be evaluated through the satellite orbit and clock determination process. Results from both approaches are consistent with each other and show that the uncertainties of the measured satellite clock offsets are less than 2 ns.

In order to transmit correct navigation signals, the clock offsets are controlled within 1 ms. Between adjustments, the offset exhibits a dominant linear trend over time (i. e., a frequency offset) and a superimposed quadratic variation (i. e., a frequency drift), which is common for Rubidium atomic frequency standards. By means of an example, Fig. 10.30 shows clock offset variation of a geostationary BeiDou satellite after lin-

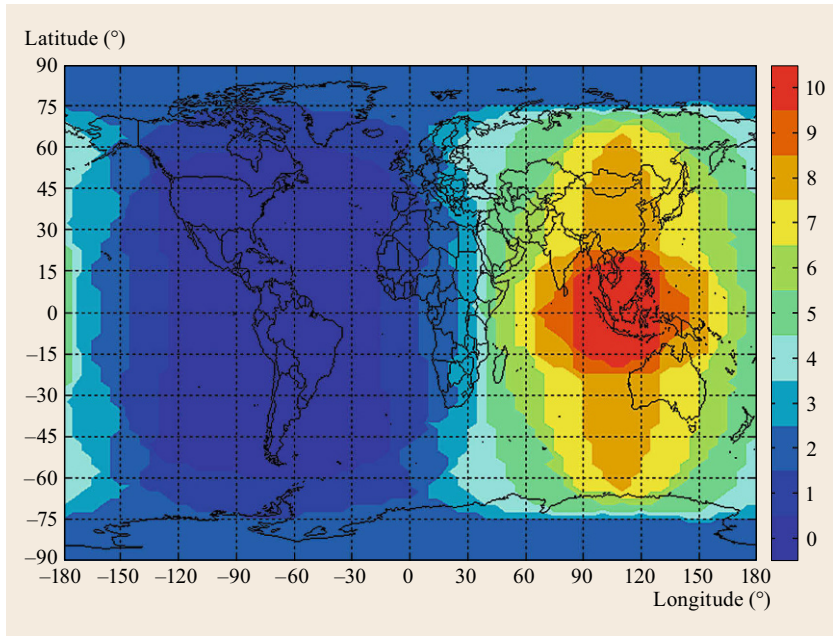


Fig. 10.27 Number of visible BeiDou satellites (95%) in January 2013 considering a cut-off elevation of 5° (after [10.51])

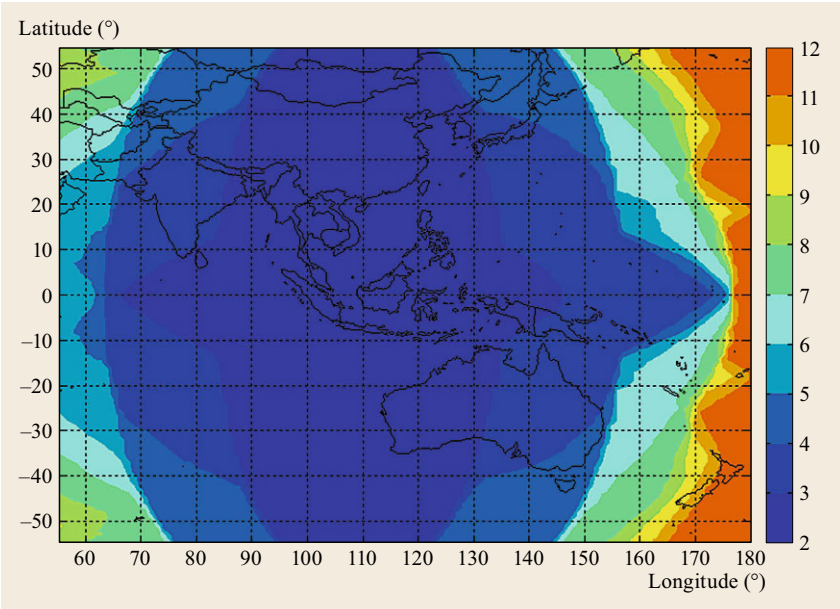


Fig. 10.28 PDOP value (95%) in Asia-Pacific area (after [10.51])

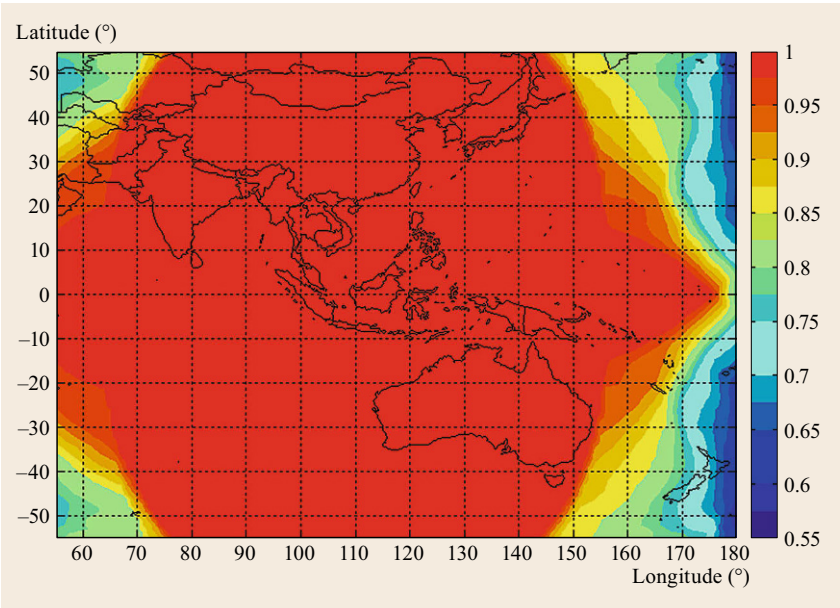


Fig. 10.29 Availability of BeiDou in Asia-Pacific area (PDOP < 6) (after [10.51])

ear and quadratic detrending over a 2-month interval in 2014.

Frequency offsets, drifts, and stabilities (i.e., ADEV for 1-day intervals) as obtained from TWTT [10.54–56] between the satellites and the seven type-A monitoring stations are summarized in Table 10.7 for all BeiDou-2 satellites. The satellite clocks (SCs) from three different companies are symbolized as SC1, SC2, SC3, and SC4, respectively.

As can be seen, the performances of the four clocks are comparable, only the accuracy of SC1 is slightly lower.

The frequency stability as obtained from the TWTT observations over time scales from 1 s to several days is illustrated in Fig. 10.31. Here, ADEV values of $2\text{--}8 \cdot 10^{-14}$ over one day are obtained for most satellites, while a stability of about $5 \cdot 10^{-11}$ is obtained near 1 s. However, the measured ADEV is dominated by the noise of the employed ranging measurements and

Table 10.7 Performance of BDS satellite clocks in April 2013. Satellites are identified by their type (*G*, *I*, and *M* for GEO, IGSO, and MEO satellites, respectively) as well as the PRN number of the transmitted signal. The last column indicates, which of the four clocks of each satellite was activated during the analysis period

Sat	PRN	Frequency offset (10^{-11})	Drift (10^{-13} /d)	Stability (at 1 day) (10^{-13})	Clock
G1	C01	+1.935	-0.351	0.741	SC3
G6	C02	-4.147	+2.299	3.671	SC2
G3	C03	+0.513	+0.272	0.385	SC2
G4	C04	+0.769	+0.279	1.123	SC4
G5	C05	-5.271	-1.450	0.704	SC1
I1	C06	-1.208	+0.060	1.792	SC1
I2	C07	+1.158	+0.331	1.359	SC3
I3	C08	+0.890	+0.698	2.698	SC4
I4	C09	+2.778	-0.352	0.493	SC1
I5	C10	+2.929	-0.752	0.702	SC3
M3	C11	+0.144	-2.448	0.508	SC1
M4	C12	+1.047	-0.445	0.187	SC2
M5	C13	+5.969	+3.679	1.116	SC1
M6	C14	+0.858	+5.752	2.412	SC2

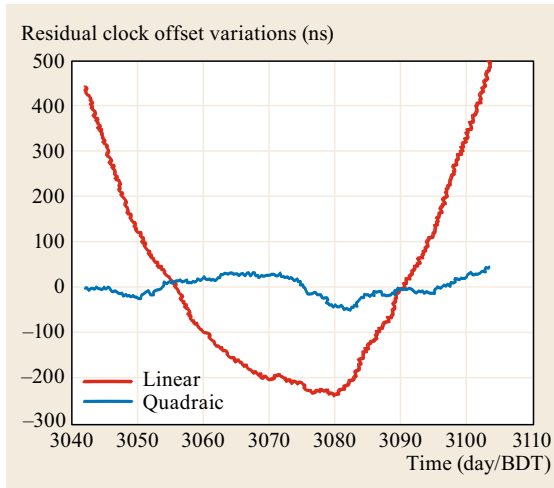


Fig. 10.30 Residual clock offset variations of GEO-3 satellite clock after linear and quadratic detrending from May 1 to July 1, 2014

does not reflect the actual onboard clock performance below an averaging time of 1000 s [10.55]. Independent analyses of one-way carrier phase observations reported in [10.57] and [10.58] have actually demonstrated a short-term stability of about $0.5 \cdot 10^{-11}$ at 1 s intervals. Modified ADEVs of less than $1 \cdot 10^{-14}$ at daily time scales are reported in [10.59].

10.3.3 Positioning Performance

The BDS-2 positioning performance following the start of the regional service has been assessed in [10.51].

Single-point position solutions were obtained from measurements collected with a geodetic receiver in Beijing from December 27, 2012 to March 20, 2013 (including a 13-day unavailability due to receiver update and maintenance). The employed B II pseudorange observations exhibit a noise level of 0.2–0.4 m and the BeiDou Klobuchar model was used for compensating ionospheric path delays. Observations were limited to a 5° cut-off elevation.

The resulting 95% (2σ) position errors relative to the surveyed antenna location are illustrated in Fig. 10.32 over the 71-day time interval. Complementary to these, the horizontal, vertical, and PDOP (95% values) of every day are shown in Fig. 10.33. On most days, the 95th percentiles of the horizontal, vertical, and total position error are smaller than 6 m, 10 m, and 12 m, respectively, which is in good accord with the open service performance specification [10.10].

The average single-point positioning accuracy of 10 m at a mean PDOP of 3 indicates a users range error (URE) of about $3.5 \text{ m } 2\sigma$. This value includes user equipment related contributions (noise, multipath, and uncompensated atmospheric delays) as well as the SISRE, which characterizes the contribution of broadcast orbit and clock errors. Independent comparisons of broadcast ephemerides with postprocessed precise orbit and clock determination results indicate a SISRE value of about 1.5 m (rms) on average over the entire BDS-2 constellations [10.60–62].

Apart from standard positioning applications, BeiDou has found widespread use for precise point positioning in stand-alone or multiconstellation mode. Precise orbit and clock products for BeiDou are rou-

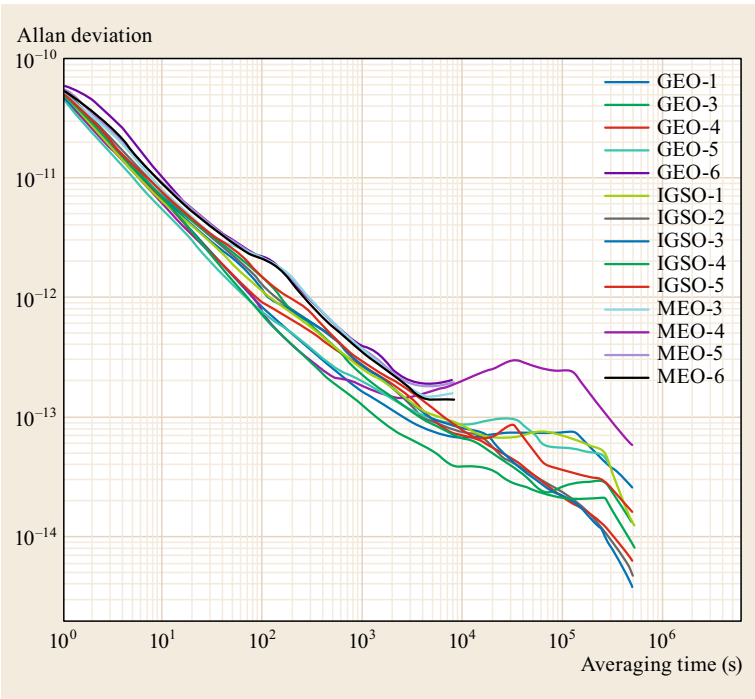


Fig. 10.31 Frequency stability (ADEV) of BeiDou satellite clocks based on two-way time transfer (after [10.55])

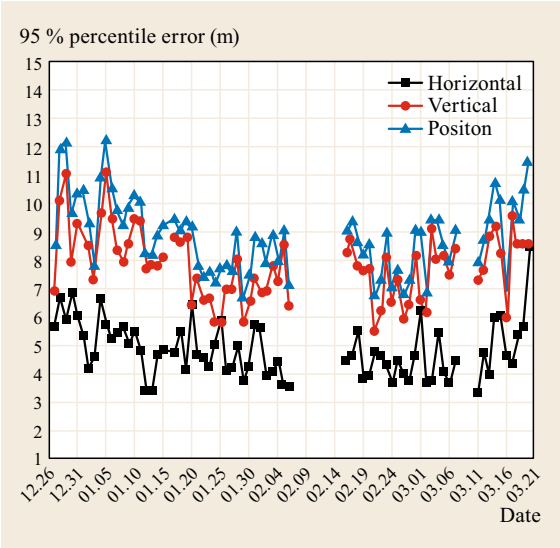


Fig. 10.32 Positioning performance of BeiDou-only navigation in Beijing after start of the BDS-2 regional service from December 2012 to March 2013 (after [10.51])

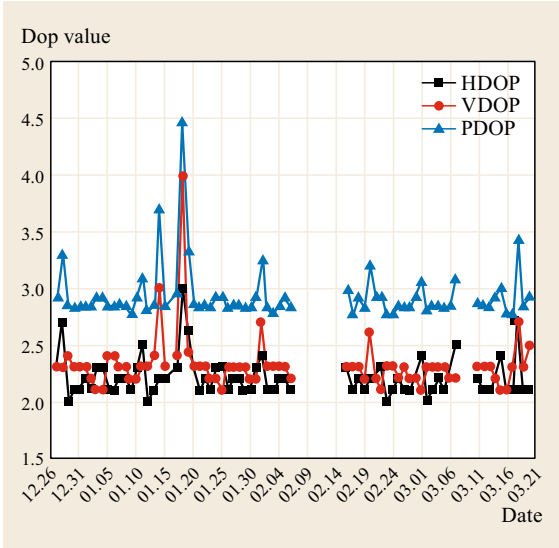


Fig. 10.33 Horizontal (HDOP), vertical (HDOP) and PDOP of BDS-2 in Beijing between December 2012 and March 2013 (after [10.51])

tinely generated by a variety of analysis centers [10.59, 63, 64] and publicly made available through the International GNSS Service.

A comprehensive assessment of the BeiDou tracking performance and the associated positioning performance conducted in [10.51] is summarized in

Fig. 10.34. Using a geodetic grade receiver, code and carrier phase observations with noise levels of about 10 cm and 0.5 mm, respectively, were collected. While the code-only position exhibit rms errors of 6 m and 10 m in horizontal (H) and up (U) directions, respectively, the cancellation of common errors in differential

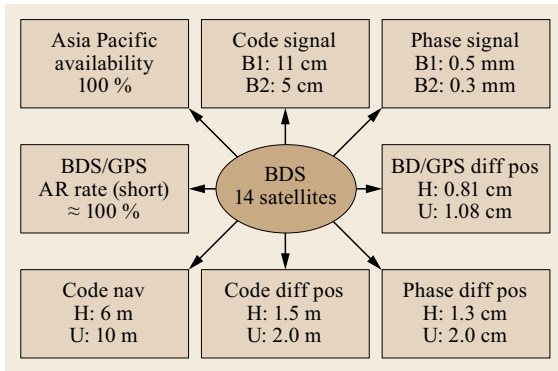


Fig. 10.34 Performance of BDS-2 service in Beijing area (after [10.51]). See text for further explanations

code positioning enables an accuracy better than 2 m in each direction. Using carrier phase observations, differential positions with centimeter-level accuracy and an ambiguity resolution (AR) rate of about 100% are obtained at short baselines. Combining BDS and GPS, the carrier phase-based differential relative navigation accuracy can further be improved by a factor of 2.

10.3.4 Application Examples

Starting with the successful implementation of the BeiDou Navigation Satellite Demonstration System in 2003, BDS has been used for a wide range of applications and created substantial social and economic benefits [10.4]. BeiDou terminals and receivers developed by Chinese industry are employed for land and maritime navigation, provide precise timing for critical infrastructure, contribute to weather services, and are widely employed for emergency services. Prominent examples highlighting the benefits of BeiDou for the Chinese society include the use of BDS navigation during the Beijing Olympic Games and the Shanghai World Expo as well as as BDS navigation and communication in earthquake and snow disaster relief.

Use of the BeiDou Navigation Satellite System in the transportation sector has been promoted through pilot projects such as the Demonstration System of Monitoring Management Services in Priority Transportation, the Highway Infrastructure Safety Monitoring System [10.65, 66], and School Bus Safety Mon-

itoring [10.67]. In the maritime sector, an integrated information service platform has been established to provide vessel localization and monitoring, rescue and information services as well as harbor access management for fishing boats [10.4, 68].

Apart from the navigation related applications, BeiDou is an important element for the reliable distribution of precise time information. Key technologies such as long-distance fiber technology and an integrated satellite-based timing system have been developed as part of the BeiDou two-way timing demonstration program [10.69, 70]. Similarly, a pilot project for power system time synchronization provided the basis for monitoring and protection of electric power grids [10.71].

The pilot project for *Monitoring and Warning in Atmospheric, Oceanic and Space* has addressed the automatic data transmission among China's meteorological stations and a real-time hydrological monitoring system has been established to support flood and drought control [10.4]. Further studies in this field include the assessment of precipitable water vapor from ground-based BeiDou observations in a precise point positioning approach [10.72] as well as the estimation of typhoon wind speeds using reflected signals from BeiDou GEO satellites and the transfer of related information using the BeiDou short message mode [10.73].

BeiDou also plays a vital role in diverse forms of emergency and disaster relief services. Among others, forest fire monitoring terminals based on BDS have been designed and produced by Chinese industry [10.74]. They have been successfully used in a forest fire prevention system [10.75], which benefits from the combination of the BeiDou positioning and short message communication services. This unique feature of BeiDou is also utilized within the nationwide disaster relief management system, where BeiDou has greatly improved the efficiency of emergency operations and decision-making processes [10.4, 76, 77].

Finally, with the ongoing proliferation of navigation satellite systems, BeiDou has become an important contributor to multi-GNSS. With the availability of more satellites and signals, precision and reliability of parameter estimation improves, as well as convergence times, position availability and robustness of ambiguity resolution [10.78–80].

10.4 BeiDou (Global) Navigation Satellite System

The BeiDou Navigation Satellite System with global coverage (BDS-3) will be completed by 2020. The space constellation will consist of five GEO satellites

(positioned at 58.75°E, 80°E, 110.5°E, 140°E, and 160°E) as well as as 27 MEO satellites and three IGSO satellites. The current satellites will be a part of the

constellation that forms the BeiDou global navigation satellite system. Future MEO and IGSO satellites will have the same orbit as the current ones, that is, they will be placed in orbits with an inclination of 55° at altitudes of 21 500 and 36 000 km, respectively. The three IGSO satellites orbit the Earth in three different planes but exhibit a common ground track with its ascending node at 118°E .

BDS-3 will provide an open service and an authorized service in four frequency bands, including B1 (1559–1610 MHz), B2 (1164–1219 MHz), and B3 (1240–1300 MHz) with center frequencies of 1575.42 MHz, 1191.795 MHz, and 1268.52 MHz. A new S-band signal, Bs (2483.5–2500 MHz), is also broadcasted by the newly launched satellites. Compared to the regional BeiDou system, new and advanced signal structures are employed for better performance, compatibility, and interoperability with other GNSSs [10.81]. Key characteristics of these signals as presented in [10.82–84] are summarized in the following.

The B1 frequency has two signals, named B1-A and B1-C. B1-A is an authorized signal and employs a BOC(14,2) binary offset carrier modulation. The open service B1-C signal, in contrast, utilizes a MBOC(6,1,1/11) multiplexed binary offset carrier modulation. It consists of two components, namely the B1-CD data channel and the B1-CP pilot channel, which are transmitted in phase quadrature. The B1-CD component is modulated with a 50 bps (100 sps) binary navigation data stream.

On the B2 frequency, an AltBOC(15,10) modulation is used, which generates two side lobes at 1176.45 MHz (B2a) and 1207.14 MHz (B2b), respectively, and comprises four individual signal components

for the open service. B2a consists of a data channel (B2a-D) and a pilot channel (B2a-P) in phase quadrature. The binary navigation data stream on B2a-D is transmitted at a rate of 25 bps (50 sps). Similarly, the upper sideband comprises the B2b-D data channel (with twice the data rate of the B2a-D channel) and a B2b-P pilot channel in phase quadrature.

In the B3-band, two authorized signals (B3 and B3-A) with a total of four individual components are transmitted at a common center frequency of 1268.52 MHz. The B3I and B3Q signals are modulated in quadrature phase shift keying (QPSK) with a 10.23 Mchips/s PRN code and a 500 bps binary navigation data stream. The B3-A signal is modulated in BOC(15,2.5) and consists of two components, B3-AD and B3-AP, in phase quadrature. The B3-AD data channel is modulated with a 50 bps (100 sps) binary navigation message.

The Bs signal, finally, is modulated in BPSK(8) and transmitted at a central frequency of 2492.028 MHz. It consists of two components, designated as Bs-D and Bs-P. The Bs-D component is modulated with a 50 bps (100 sps) binary navigation data stream.

Between March and September 2015, the first four satellites of the third generation BeiDou system (including two IGSO and two MEO satellites) have been launched. Initial test signals transmitted by these satellites for system validation purposes are described in [10.85]. As of December 2015, the final structure of the BDS-3 signals has not officially been published and minor modifications of the aforementioned signal properties might still be performed for further optimization. However, as China had claimed before, the ICD of the global open signal will be published as soon as possible once the new signal system assessment is accomplished [10.81].

10.5 Brief Introduction of CAPS

The Chinese Area Positioning System (CAPS) is an alternative, regional radio navigation system that has been developed since the early 2000s under the lead of the Chinese National Astronomical Observatories and the NTSC. It implements the concept of a *transponder satellite communication navigation and positioning system* [10.86] utilizing spare capacities of existing geosynchronous communication satellites. While traditional GNSS concepts are build on a dedicated satellite constellation, which generates and transmits the navigation signals in orbit, CAPS makes use of communication satellites, which relay navigation signals generated on ground to the users. The reuse of existing space infrastructure offers considerable cost savings

along with great flexibility and redundancy [10.87–89].

CAPS has been involved in the testing of signals, time synchronization, and orbit determination of the BeiDou system. In the future, CAPS will contribute to the performance monitoring of BeiDou and serve as an experimental platform for the integration of navigation and communication services.

10.5.1 CAPS Concept and System Architecture

Similar to other spacebased navigation systems, the CAPS system can be divided into a space segment and a ground segment (Fig. 10.35).

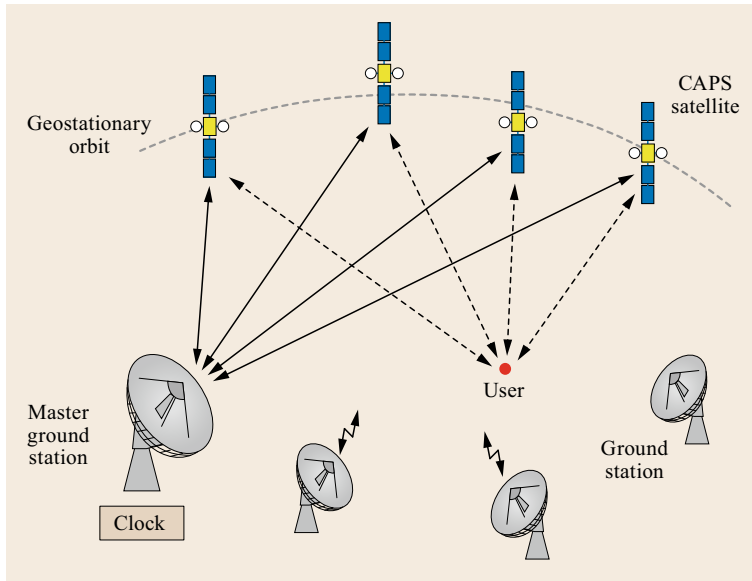


Fig. 10.35 The CAPS system architecture comprises various satellite in near-geostationary orbit as well as a master ground station and multiple remote stations for orbit determination

The space segment is made up of multiple communications satellites in geosynchronous orbits, which offer continued visibility from the regional service area. Apart from satellites in a tightly controlled GEO, the CAPS includes decommissioned GEO (DGEO) satellites which no longer perform a north-south station keeping and attain a slightly inclined geosynchronous orbit (SIGSO). The use of highly inclined geosynchronous orbits has, furthermore, been suggested to improve the geometric distribution of visible satellites and the overall dilution of precision (DOP) in CAPS-based positioning [10.87]. The satellites do not require a dedicated navigation payload and high performance atomic clock, but merely a standard transponder to relay a signal between a ground based transmitter and a receiver. As such, CAPS offers great flexibility in the choice of satellites for its space segment.

The CAPS ground segment consists of a master ground station and various remote ground stations. All stations are equipped with large dish antennas for transmitting a modulated ranging signal to the satellites and can simultaneously receive signals from other stations returned through the satellites' transponders. The master station hosts the primary time and frequency reference. It generates the CAPS system time, which itself is synchronized to the realization of Universal Coordinated Time provided by the National Time Service Center (UTC(NTSC)) [10.90].

The master station also serves as the ground navigation center, which determines the orbits and transponder delays of the CAPS satellites from the measured measures transmission times between stations, and broadcasts this information to the user as part of the naviga-

tion message. In addition to orbit and clock information, the CAPS navigation message includes near real-time temperature and pressure data collected by meteorological stations across China. Along with barometers in the user terminals, this information can be used to infer the altitude of the user and to support a robust three-dimensional positioning.

Users receiving ranging signals from multiple CAPS satellites, can determine their position using the measured signal travel times, and, optionally, barometric height information. Even though, at first sight, CAPS resembles the first-generation BeiDou system (Sect. 10.1), it employs a completely passive navigation technique. No signals need to be returned by the user to the control center and all positions are computed in the end-user equipment rather than a central navigation facility.

CAPS utilizes C-band links for all signals exchanged between ground stations, satellites and users. This choice is partly driven by the wide availability of C-band transponder capacity on geostationary communication satellites, but also motivated by a general interest to avoid the crowded L-band spectrum for new navigation systems. As discussed in [10.91], the higher frequency induces increased free-space losses and a higher attenuation due to rainfall, but is substantially less affected by ionospheric path delays. Also, C-band supports the use of wide-band navigation signals with beneficial noise and multipath properties.

To avoid interference at the satellites, different frequencies need to be employed for the uplink and downlink. For ionospheric correction, the satellites of the CAPS trial system described in Sect. 10.5.3

transmit dual-frequency navigation signals at $f_{C1} = 4143.15$ MHz and $f_{C2} = 3826.02$ MHz to its users. The ground stations, in contrast, use carrier frequencies of $f_{C1} = 6368.15$ MHz and $f_{C2} = 6051.02$ MHz for the uplink. The uplink and downlink signals are separated by a constant offset of 2225 MHz, which is generated by a local oscillator onboard the satellites [10.92]. Since the onboard oscillator exhibits only a moderate stability, an active alignment of the uplink frequency is performed in the master station to compensate the frequency shift of the transponder. In this way, CAPS can also be used to determine the user velocity based on observed Doppler shifts at the receiver.

Through the inherent combination of navigation and communication, CAPS offers unique advantages over traditional navigation-only satellite systems. Possible applications of these features include the provision of differential corrections and integrity information but also various forms of search and rescue or disaster management services. The design and application of a CAPS user-terminal supporting joint navigation and communication functions for land and marine use is, for example, discussed in [10.93].

10.5.2 Positioning Principle of CAPS

Navigation as well as orbit determination and time synchronization in the CAPS are performed through measurements of signal travel times between the transmitting ground station and the user receiver. The signal path involves an uplink from the ground station at \mathbf{r}_g to the transponding satellite at \mathbf{r}_s and a subsequent downlink to the user at \mathbf{r}_u . Denoting the transmit and receive times by t_g and t_u , respectively, the two-leg pseudorange is given by

$$\begin{aligned} p_{gsu} &= c \cdot (t_u + dt_u - t_g) \\ &= ||\mathbf{r}_s - \mathbf{r}_g|| + ||\mathbf{r}_u - \mathbf{r}_s|| \\ &\quad + c\tau_s + cdt_u + T_{gs} + T_{su} . \end{aligned} \quad (10.9)$$

Here, τ_s is the satellite transponder delay and dt_u denotes the offset of the receiver clock from CAPS system time, which is defined by the master ground station. The tropospheric delays of the uplink and downlink paths are denoted by T_{gs} and T_{su} , respectively. They can be taken into account through meteorological measurements at the ground station and tropospheric models for the receiver location. Ionospheric delays are eliminated through the use of a ionosphere-free dual-frequency combination and have, therefore, been ignored in the pseudorange model.

Equation (10.9) describes the basic relation among the measured signal travel times, the satellite, and sta-

tion coordinates as well as the user position and clock offset. Depending on the specific application, different alternative formulations can be employed. First, the relations

$$\begin{aligned} p_{gsg} &= 2||\mathbf{r}_s - \mathbf{r}_g|| + c\tau_s + 2T_{gs} \\ p_{gsg'} &= ||\mathbf{r}_s - \mathbf{r}_g|| + ||\mathbf{r}_{g'} - \mathbf{r}_s|| \\ &\quad + c\tau_s + cdt_{g'} + T_{gs} + T_{sg'} \end{aligned} \quad (10.10)$$

are obtained for pseudorange observations collected by the master ground station as well as pseudorange observations between the master station (g) and the remote ground stations (g'). These can essentially be used to measure the instantaneous distance between each station and the satellite, and also to determine the three-dimensional satellite position. They also enable calibration of the satellite transponder delay and (by comparison of $p_{gsg'}$ and $p_{g'sg}$) a clock synchronization of the remote stations with the master ground station.

Second, (10.9) may be rephrased as

$$\begin{aligned} p_{su} &= p_{gsu} - c \cdot \tau_{VCLK} \\ &= ||\mathbf{r}_u - \mathbf{r}_s|| + cdt_u + T_{su} . \end{aligned} \quad (10.11)$$

Within this *virtual clock* (VCLK) concept [10.94]

$$\begin{aligned} c \cdot \tau_{VCLK} &= c \cdot (t_s - t_g) \\ &= ||\mathbf{r}_s - \mathbf{r}_g|| + c\tau_s + T_{gs} . \end{aligned} \quad (10.12)$$

denotes the difference between the uplink time t_g at the ground station and the time t_s of retransmission at the satellite. The VCLK correction is determined at the ground station from the known satellite orbit and calibrated time delays. It is then transmitted to the user via the CAPS navigation message and can be used to remove the uplink contribution from the measured pseudorange. As a result, the measurement model (10.11) attains the well-known form of traditional navigation satellite systems using one-way pseudorange observations. The user position and receiver clock offset can then be determined from a minimum of four corrected pseudorange observations. While a closed-form solution is possible for exactly four observations, the measurement model is commonly linearized with respect to the user coordinates and solved in an iterative manner.

Due to the regional distribution of the ground stations as well as the large distance and small orbital inclinations of the near-geostationary satellites, CAPS suffers from an unfavorable dilution of precision (DOP) as compared to traditional GNSSs. The use of barometric measurements is, therefore, considered as a pseudo-observation to better constrain the user position [10.95].

Making use of a barometer in the user terminal as well as broadcast meteorological data (ground pressure and temperature) for the region of interest, the receiver's height above the reference ellipsoid can be determined with an accuracy ranging from 3–5 m for terrestrial users to 10–20 m for aviation users [10.90].

10.5.3 Trial CAPS System

Following the governmental approval in 2005, a trial system was established in the course of two years at a cost of about 20 million USD to validate the working principle and performance of the CAPS.

The trial constellation consists of two GEO satellites (Zhongwei/ChinaStar-1 and Sinosat-1) located at 87.5°E and 110.5°E, as well as two decommissioned GEOs (Apstar-1A and Apstar-1), located at 130°E and 142°E, respectively. The master ground station of the CAPS prototype is located at Lintong in central China and hosts a total of six 7 m antennas for uplink and downlink to/from the geosynchronous satellites [10.89]. Four additional ground stations at Urumqi, Shanghai, Changchun, and Kunming provide the necessary measurements for orbit determination.

Each of the two carriers (C1 and C2) is modulated with two ranging signals namely a coarse-and-acquisition (C/A) with a 1.023 MHz chipping rate and an encrypted precise (P) code at 10.23 MHz [10.90, 96]. The CAPS navigation message is transmitted at a rate of 50 bps and comprises 44 frames with a length of 1500 bits.

Static and dynamic performance tests were performed in various regions throughout the China mainland. In addition to the actual navigation signals, barometric altimetry is employed as a virtual satellite. Following [10.87], horizontal positioning accuracies of 15–25 m and 5–10 m were achieved using the coarse and precise ranging signals, respectively, while the velocity accuracy attained values in the range of 0.1–0.3 m/s.

As of 2015, this system is used as a preliminary operational system, and a complete system with three GEO, three DGEO, and three IGSO is pursued.

Acknowledgments. The authors are indebted to Dr. Li LIU, Dr. Zhiwu CAI, Ms. Jinxian ZHAO, and Ms. Xia GE of Beijing Satellite Navigation Center for their valuable assistance.

References

- 10.1 J. Needham, W. Ling, K.G. Robinson: *Science and Civilisation in China, Vol. 4: Physics and Physical Technology* (Cambridge Univ. Press, Cambridge 1962) pp. 239–278
- 10.2 C. Ran: Development of the BeiDou Navigation Satellite System. Global navigation satellite systems, Rep. Jt. Work. Natl. Acad. Eng. Chin. Acad. Eng., Shanghai, ed. by L.A. Davis, P.K. Enge, G.X. Gao (National Academies Press, Washington 2012) pp. 83–94
- 10.3 F. Hirth: Origin of the mariner's compass in China, *Monist* **16**(3), 321–330 (1906)
- 10.4 China Satellite Navigation Office: Report on the development of BeiDou Navigation Satellite System, Version 2.2 (China Satellite Navigation Office, Beijing 2013)
- 10.5 S. Bian, J. Jin, Z. Fang: The Beidou satellite positioning system and its positioning accuracy, *Navigation* **52**(3), 123–129 (2005)
- 10.6 M.A. Rothblatt: *Radiodetermination Satellite Services and Standard* (Artech House, Norwood 1987)
- 10.7 R.D. Briskman: Radio determination satellite service, *Proc. IEEE* **78**(7), 1096–1106 (1990)
- 10.8 C. Han, Y. Yang, Z. Cai: BeiDou Navigation Satellite System and its time scales, *Metrologia* **48**(4), S213–S218 (2011)
- 10.9 J. Wei, D. Xu, J. Deng, P. Huang: Synchronization for BeiDou satellite terrestrial improvement radio navigation system, *Int. Conf. Intell. Mechatron. Automation*, Chengdu (2004) pp. 672–676
- 10.10 BeiDou Navigation Satellite System open service performance standard, Version 1.0 (China Satellite Navigation Office, Beijing 2013)
- 10.11 BeiDou Navigation Satellite System signal in space interface control document – Open service signal, Version 2.0 (China Satellite Navigation Office, Beijing 2013)
- 10.12 J.G. Walker: Satellite constellations, *J. Br. Interplanet. Soc.* **37**, 559–572 (1984)
- 10.13 P. Steigenberger, U. Hugentobler, A. Hauschild, O. Montenbruck: Orbit and clock analysis of Compass GEO and IGSO satellites, *J. Geod.* **87**(6), 515–525 (2013)
- 10.14 J. Xie, J. Wang, H. Mi: Analysis of Beidou navigation satellites in-orbit state, *Proc. Chin. Satell. Navig. Conf. (CSNC)*, Guangzhou, Vol. I, ed. by J. Sun, J. Liu, Y. Yang, S. Fan (Springer, Berlin 2012) pp. 111–122
- 10.15 L. Fan, C. Jiang, M. Hu: Ground track maintenance for BeiDou IGSO satellites subject to tesseral resonances and the luni-solar perturbations, *Adv. Space Res.* (2016), doi:[10.1016/j.asr.2016.09.014](https://doi.org/10.1016/j.asr.2016.09.014)
- 10.16 F. Neuman, L. Hofman: New pulse sequences with desirable correlation properties, *Proc. Natl. Telem. Conf.*, Washington (1971) pp. 272–282
- 10.17 D. Zou, Z. Deng, J. Huang, H. Liu, L. Yang: A study of Neuman Hoffman codes for GNSS application, *Proc. 5th Int. Conf. Wirel. Commun. Netw. Mob. Comput.* Beijing (2009) pp. 1–4

- 10.18 C.J. Hegarty: GNSS signals – An overview, IEEE Int. Conf. Freq. Cont. Symp. (FCS) (2012) pp. 1–7
- 10.19 N. Nadarajah, P.J.G. Teunissen, J.-M. Sleewaegen, O. Montenbruck: The mixed-receiver BeiDou inter-satellite-type bias and its impact on RTK positioning, GPS Solutions **19**(3), 357–368 (2015)
- 10.20 Z. Li, H. Wu, L. Wang, H. Liu: Research on the BDS inter-satellite-type carrier phase bias introduced by different NH code sign conventions, Proc. Chin. Satell. Navig. Conf. (CSNC), Xi'an, Vol. I, ed. by J. Sun, J. Liu, S. Fan, X. Lu (Springer, Berlin 2015) pp. 805–816
- 10.21 M. Shi, A. Peng, G. Ou: Analysis to the effects of NH code for Beidou MEO/IGSO satellite signal acquisition, IEEE 9th Conf. Ind. Electron. Appl. (ICIEA), Hangzhou (2014) pp. 2075–2080
- 10.22 M.Z.H. Bhuiyan, S. Söderholm, S. Thombre, J. Ruotsalainen, H. Kuusniemi: Overcoming the challenges of BeiDou receiver implementation, Sensors **14**(11), 22082–22098 (2014)
- 10.23 T. Grelier, J. Dantepal, A. Delatour, A. Ghion, L. Ries: Initial observations and analysis of compass MEO satellite signals, Inside GNSS **2**(4), 39–43 (2007)
- 10.24 G.X. Gao, A. Chen, S. Lo, D. De Lorenzo, T. Walter, P. Enge: Compass-M1 broadcast codes in E2, E5b, and E6 frequency bands, IEEE J. Sel. Top. Sig. Process. **3**(4), 599–612 (2009)
- 10.25 W. Tang, C. Deng, C. Shi, J. Liu: Triple-frequency carrier ambiguity resolution for Beidou navigation satellite system, GPS Solutions **18**(3), 335–344 (2014)
- 10.26 J. Li, Y. Yang, J. Xu, H. He, H. Guo: GNSS multi-carrier fast partial ambiguity resolution strategy tested with real BDS/GPS dual-and triple-frequency observations, GPS Solutions **19**(1), 5–13 (2015)
- 10.27 N. Nadarajah, P.J.G. Teunissen, N. Raziq: Instantaneous BeiDou-GPS attitude determination: A performance analysis, Adv. Space Res. **54**(5), 851–862 (2014)
- 10.28 P.J.G. Teunissen, R. Odolinski, D. Odijk: Instantaneous BeiDou+GPS RTK positioning with high cut-off elevation angles, J. Geod. **88**(4), 335–350 (2014)
- 10.29 R.C. Bose, D.K. Ray-Chaudhuri: On a class of error correcting binary group codes, Inf. Control **3**(1), 68–79 (1960)
- 10.30 O. Montenbruck, P. Steigenberger: The BeiDou navigation message, J. Glob. Position. Syst. **12**(1), 1–12 (2013)
- 10.31 F. Guo, X. Zhang, J. Wang: Timing group delay and differential code bias corrections for BeiDou positioning, J. Geod. **89**, 427–445 (2015)
- 10.32 X. Wu, X. Hu, G. Wang, H. Zhong, C. Tang: Evaluation of COMPASS ionospheric model in GNSS positioning, Adv. Space Res. **51**(6), 959–968 (2013)
- 10.33 J. Xie, T. Liu: Research on technical development of BeiDou navigation satellite system, Proc. Chin. Satell. Navig. Conf. (CSNC), Wuhan, Vol. I, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin 2013) pp. 197–209
- 10.34 A. Gilks: China's space policy: Review and prospects, Space Policy **13**(3), 215–227 (1997)
- 10.35 W. Wang, G. Chen, S. Guo, X. Song, Q. Zhao: A study on the Beidou IGSO/MEO satellite orbit determination and prediction of the different yaw control mode, Proc. Chin. Satell. Navig. Conf. (CSNC), Wuhan, Vol. III, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin 2013) pp. 31–40
- 10.36 J. Guo, Q. Zhao, T. Geng, X. Su, J. Liu: Precise orbit determination for COMPASS IGSO satellites during yaw maneuvers, Proc. Chin. Satell. Navig. Conf. (CSNC), Wuhan, Vol. III, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin 2013) pp. 41–53
- 10.37 J. Guo, Q. Zhao: Analysis of precise orbit determination for BeiDou satellites during yaw maneuvers, Proc. Chin. Satell. Navig. Conf. (CSNC), Wuhan (2014)
- 10.38 S. Zhou, X. Hu, J. Zhou, J. Chen, X. Gong, C. Tang, B. Wu, L. Liu, R. Guo, F. He, X. Li, H. Tan: Accuracy analyses of precise orbit determination and timing for COMPASS/Beidou-2 4GEO/5IGSO/4MEO constellation, Proc. Chin. Satell. Navig. Conf. (CSNC), Wuhan, Vol. III, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin 2013) pp. 89–102
- 10.39 L.A. Mallette, J. White, P. Rochat: Pace qualified frequency sources (clocks) for current and future GNSS applications, IEEE/ION PLANS, Indian Wells (2010) pp. 903–908
- 10.40 J. Lu: COMPASS/Beidou navigation satellite system development, 3rd Meet. Int. Comm. GNSS (ICG), Pasadena (UNOOSA, Vienna 2008) pp. 1–42
- 10.41 Z.-P. Zhang, H.-F. Zhang, W.-Z. Chen, P. Li, W.-D. Meng, Y.-M. Wang, J. Wang, W. Hu, F.-M. Yang: Design and performances of laser retro-reflector arrays for Beidou navigation satellites and SLR observations, Adv. Space Res. **54**(5), 811–817 (2014)
- 10.42 W. Meng, H. Zhang, P. Huang, J. Wang, Z. Zhang, Y. Liao, Y. Ye, W. Hu, Y. Wang, W. Chen, F. Yang, I. Prochazka: Design and experiment of onboard laser time transfer in Chinese Beidou navigation satellites, Adv. Space Res. **51**(6), 951–958 (2013)
- 10.43 I. Prochazka, F. Yang: Photon counting module for laser time transfer via Earth orbiting satellite, J. Mod. Opt. **56**(2/3), 253–260 (2009)
- 10.44 W. Song, J. Shen: China – Development of BeiDou Navigation Satellite System (BDS) – A Program update, Proc. ION Pacific PNT, Honolulu (ION, Virginia 2015)
- 10.45 Y. Yang: Chinese geodetic coordinate system 2000, Chin. Sci. Bull. **54**(15), 2714–2721 (2009)
- 10.46 D. D. McCarthy: *IERS Conventions (1996)*, IERS Technical Note No. 21, (Observatoire de Paris, Paris 1996)
- 10.47 G. Petit, B. Luzum: *IERS Conventions (2010)* IERS Technical Note No. 36, (Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt 2010)
- 10.48 Y. Yang, Y. Wen, J. Xiong, J. Yang: Robust estimation for a dynamical model of the sea surface, Surv. Rev. **35**, 2–10 (1999)
- 10.49 Y. Yang, L. Song, T. Xu: Robust estimator for correlated observations based on bifactor equivalent weights, J. Geod. **76**(6/7), 353–358 (2002)
- 10.50 C. Han, S. Xiao, Z. Cai: Progress of BDT and its relationship with UTC/UTCr, 9th Meet. Int. Comm. GNSS (ICG), Work. Group A, Prague (UNOOSA, Vienna 2014) pp. 1–22
- 10.51 Y. Yang, J.L. Li, A.B. Wang, J.X. Xu, H.B. He, H.R. Guo, J.F. Shen, X. Dai: Preliminary assessment of the

- navigation and positioning performance of BeiDou regional navigation satellite system, *Sci. Chin. Earth Sci.* **57**(1), 144–152 (2014)
- 10.52 R.B. Langley: Dilution of precision, *GPS World* **10**(5), 52–59 (1999)
- 10.53 Y. Yang, J.L. Li, J.Y. Xu, J. Tang, H.R. Guo, H.B. He: Contribution of the COMPASS satellite navigation system to global PNT users, *Chin. Sci. Bull.* **56**(26), 2813–2819 (2011)
- 10.54 L. Liu, L.F. Zhu, C.H. Han, X.P. Liu, C. Li: The model of two-way radio time transfer between the earth and satellites and analysis of its experiment, *Acta Astron. Sin.* **50**, 189–196 (2009)
- 10.55 C. Han, Z. Cai, Y. Lin, L. Liu, S. Xiao, L. Zhu, X. Wang: Time synchronization and performance of BeiDou satellite clocks in orbit, *Int. J. Navig. Obs.* **37**1450, 1–5 (2013)
- 10.56 W. Gao, Y. Lin, G. Chen, Y. Meng: The performances assessment methods and results of in-orbit atomic clocks of BDS, *J. Geomat. Sci. Technol.* **31**(4), 15–19 (2014), in Chinese
- 10.57 A. Hauschild, O. Montenbruck, P. Steigenberger: Short-term analysis of GNSS clocks, *GPS Solutions* **17**(3), 295–307 (2013)
- 10.58 E. Griggs, R. Kursinski, D. Akos: The accuracy of current GNSS signal sources for radio occultation missions, 8th FORMOSAT-3/COSMIC Data Users' Work., Boulder (UCAR, Boulder 2014)
- 10.59 Y. Lou, Y. Liu, C. Shi, B. Wang, X. Yao, F. Zheng: Precise orbit determination of BeiDou constellation: Method comparison, *GPS Solut.* **20**(2), 259–268 (2016)
- 10.60 Z.H. Hu, G. Chen, Q. Zhang, J. Guo, X. Su, X.T. Li, Q. Zhao, J. Liu: An initial evaluation about BDS navigation message accuracy, *Proc. Chin. Satell. Navig. Conf. (CSNC)*, Wuhan, Vol. 1, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin 2013) pp. 89–102
- 10.61 L. Chen, W. Jiao, X. Huang, C. Geng, L. Ai, L. Lu, Z. Hu: Study on signal-in-space errors calculation method and statistical characterization of BeiDou navigation satellite system, *Proc. Chin. Satell. Navig. Conf. (CSNC)*, Wuhan, Vol. 1, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin 2013) pp. 423–434
- 10.62 O. Montenbruck, P. Steigenberger, A. Hauschild: Broadcast versus precise ephemerides: A multi-GNSS perspective, *GPS Solutions* **19**(2), 321–333 (2015)
- 10.63 Q. Zhao, J. Guo, M. Li, L. Qu, Z. Hu, C. Shi, J. Liu: Initial results of precise orbit and clock determination for COMPASS navigation satellite system, *J. Geod.* **87**(5), 475–486 (2013)
- 10.64 Z. Deng, Q. Zhao, T. Springer, L. Prange, M. Uhlemann: Orbit and clock determination – BeiDou, *Proc. IGS Work. 2014*, Pasadena (IGS, Pasadena 2014) pp. 1–19
- 10.65 S. Liu, L. Hu: Application of Beidou Navigation Satellite System in logistics and transportation. Logistics: The emerging frontiers of transportation and development in China, 8th Int. Conf. Chin. Logist. Transp. Prof. (ICCLTP), Chengdu ed. by R. Liu, J. Zhang, C. Guan (ASCE, Reston 2008) pp. 1789–1794
- 10.66 R. Chen, S. Li, Z. Xu: Beidou NPS applied to monitor the structure safety health of bridge, *Int. J. Comput. Sci. Electron. Eng. (IJCSEE)* **2**(4), 192–195 (2014)
- 10.67 T. Han, X. Lu, D. Zou: Application of GNSS in school bus safety monitoring, *Proc. Chin. Satell. Navig. Conf. (CSNC)*, Guangzhou, Vol. 1, ed. by J. Sun, J. Liu, Y. Yang, S. Fan (Springer, Berlin 2012) pp. 215–223
- 10.68 Y. Lv, J. Xu, L. Xu, C. Qi: Based on BeiDou (COMPASS) build the environmental protection services system of Hainan marine fisheries production safety, *Proc. Chin. Satell. Navig. Conf. (CSNC)*, Nanjing, Vol. 1, ed. by J. Sun, W. Jiao, H. Wu, M. Lu (Springer, Berlin 2014) pp. 63–74
- 10.69 G. Tang, L. Liu, J. Cao, R. Su, X. Shi: Performance analysis for time synchronization with Compass satellite common-view, *Proc. Chin. Satell. Navig. Conf. (CSNC)*, Guangzhou, Vol. 1, ed. by J. Sun, J. Liu, Y. Yang, S. Fan (Springer, Berlin 2012) pp. 483–490
- 10.70 S. Ye: Beidou time synchronization receiver for smart grid, *Energy Procedia* **12**, 37–42 (2011)
- 10.71 Y. Wang, H. Zhao, C. Liu, Z. Chen, L. Teng, L. Lu: Applications of BeiDou satellite synchronization system in the power system, *Telecommun. Electr. Power Syst.* **32**(219), 54–57 (2011), in Chinese
- 10.72 M. Li, W. Li, C. Shi, Q. Zhao, X. Su, L. Qu, Z. Liu: Assessment of precipitable water vapor derived from ground-based BeiDou observations with Precise Point Positioning approach, *Adv. Space Res.* **55**(1), 150–162 (2015)
- 10.73 W. Li, D. Yang, F. Fabra, Y. Cao, W. Yang: Typhoon wind speed observation utilizing reflected signals from BeiDou GEO satellites, *Proc. Chin. Satell. Navig. Conf. (CSNC)*, Nanjing, Vol. 1, ed. by J. Sun, W. Jiao, H. Wu, M. Lu (Springer, Berlin 2014) pp. 191–200
- 10.74 H. Yu, L. Shi: Terminal design of forest-fire monitoring based on BeiDou satellite, *Comput. Meas. Contr.* **20**(4), 991–993 (2012) in Chinese
- 10.75 C. Hou, F. Zhang, H.F. Sun, X. Cao: Study of forest fire monitoring and commanding system based on COMPASS, *Proc. Chin. Satell. Navig. Conf. (CSNC)*, Guangzhou (CSNC, Beijing 2012), in Chinese
- 10.76 L. Xu, Y. Zhang: The system design of BeiDou alert publishing platform, *Proc. Chin. Satell. Navig. Conf. (CSNC)*, Shanghai (CSNC, Beijing 2011)
- 10.77 X. Wang: Study on disaster information collection and emergency commanding system based on BeiDou satellite technology, *J. Southwest China Norm. Univ.* **32**, 136–140 (2007)
- 10.78 X. Su, X. Zhana, M. Niu, Y. Zhang: Receiver Autonomous Integrity Monitoring (RAIM) performances of combined GPS/BeiDou/QZSS in Urban Canyon, *IEEJ Trans.* **9**, 275–281 (2014)
- 10.79 X. Li, M. Ge, X. Dai, X. Ren, M. Fritsche, J. Wickert, H. Schuh: Accuracy and reliability of multi-GNSS real-time precise positioning: GPS, Glonass, BeiDou, and Galileo, *J. Geod.* **89**(6), 607–635 (2015)
- 10.80 R. Odolinski, P.J.G. Teunissen, D. Odijk: Combined BDS, Galileo, QZSS and GPS single-frequency RTK, *GPS Solut.* **19**, 151–163 (2015)
- 10.81 C. Ran: Update on BeiDou Navigation Satellite System, 10th Meet. Int. Comm. GNSS (ICG), Boulder (UNOOSA, Vienna 2015)

- 10.82 ITU: Description of systems and networks in the radionavigation-satellite service (space-to-Earth and space-to-space) and technical characteristics of transmitting space stations operating in the bands 1164–1215 MHz, 1215–1300 MHz and 1559–1610 MHz, Recommendation M 1787, rev. 2, Sep. 2014 (ITU, Geneva 2014) <https://www.itu.int/rec/R-REC-M.1787/en>
- 10.83 S.-S. Tan, B. Zhou, S.-T. Guo, Z.-J. Liu: Research on COMPASS navigation signals of China, *Chin. Space Sci. Technol.* **31**(4), 9–14 (2011), in Chinese
- 10.84 C. Ran: BeiDou navigation satellite system development, 5th Meet. Int. Comm. GNSS (ICG), Turin (UNOOSA, Vienna 2015)
- 10.85 W. Xiao, W. Liu, G. Sun: Modernization milestone: BeiDou M2-S initial signal analysis, *GPS Solutions* **20**(2), 125–133 (2015)
- 10.86 G. X. Ai, H. L. Shi: Transponder Satellite Communication Navigation and Positioning System, PRC Patent Ser., Vol. 200410046064.1 (2004)
- 10.87 G.X. Ai, H.L. Shi, H.T. Wu, Y.H. Yan, Y.J. Bian, Y.H. Hu, Z.G. Li, J. Guo, X.D. Cai: A positioning system based on communication satellites and the Chinese Area Positioning System (CAPS), *Chin. J. Astron. Astrophys.* **8**(6), 611–630 (2008)
- 10.88 B. Li, A.G. Dempster: China's Regional Navigation Satellite System – CAPS, *Inside GNSS* **5**(4), 59–63 (2010)
- 10.89 G.Y. Ma, Q.T. Wan, T. Gan: Communication-based positioning systems: Past, present and prospects, *Res. Astron. Astrophys.* **12**(6), 601 (2012)
- 10.90 G.X. Ai, H.L. Shi, H.T. Wu, Z.G. Li, J. Guo: The principle of the positioning system based on communication satellites, *Sci. China G* **52**(3), 472–488 (2009)
- 10.91 M. Irsigler, G.W. Hein, A. Schmitz-Peiffer: Use of C-Band frequencies for satellite navigation: Benefits and drawbacks, *GPS Solutions* **8**(3), 119–139 (2004)
- 10.92 H.T. Wu, Y.J. Bian, X.C. Lu, X.H. Li, D.N. Wang: Time synchronization and carrier frequency control of CAPS navigation signals generated on the ground, *Sci. China G* **52**(3), 393–401 (2009)
- 10.93 S.M. Li, J.S. Hou, Z.R. Wang, J.T. Fan: Design of the CAPS navigation and communication incorporated terminals, *Proc. Chin. Satell. Navig. Conf. (CSNC)*, Guangzhou, Vol. III, ed. by J. Sun, J. Liu, Y. Yang, S. Fan (Springer, Berlin 2012) pp. 581–590
- 10.94 X.H. Li, H.T. Wu, Y.J. Bian, D.N. Wang: Satellite virtual atomic clock with pseudorange difference function, *Sci. China G* **52**(3), 353–359 (2009)
- 10.95 G.X. Ai, P.X. Sheng, J.L. Du, Y.G. Zheng, X.D. Cai, H.T. Wu, Y.H. Hu, Y. Hua, X.H. Li: Barometric altimetry system as virtual constellation applied in CAPS, *Sci. China G* **52**(3), 376–383 (2009)
- 10.96 Y.H. Hu, Y. Hua, L. Hou, J.F. Wei, J.F. Wu: Design and implementation of the CAPS receiver, *Sci. China G* **52**(3), 445–457 (2009)

Regional Systems

11. Regional Systems

Satoshi Kogure, A.S. Ganeshan, Oliver Montenbruck

Other than global positioning system (GPS), Russian global navigation satellite system (GLONASS), BeiDou, and Galileo, the regional navigation satellite systems (RNSS) aim to provide a regional service using a constellation of satellites in geostationary Earth orbits (GEO) and inclined geosynchronous orbits (IGSO). Two regional systems implemented in Asia will be introduced in this chapter.

The first one is the Japanese Quasi-Zenith Satellite System (QZSS), which was originally planned as an augmentation system to enhance GPS capability and performance in the area surrounding Japan. The other is the Indian Regional Navigation Satellite System (IRNSS, also known as NavIC for Navigation with Indian Constellation), which can provide an independent positioning, navigation, and timing (PNT) service over India and surrounding areas.

In this chapter, the concept of regional navigation satellite systems is first described. The combination of satellites in GEO and IGSO is a common idea to realize such a regional service platform with a low number of satellites. The orbital characteristics and geometry of the proposed RNSS constellations are explained before each RNSS is introduced in detail. Secondly, the detailed characteristics of both systems are described in the

11.1	Concept of Regional Navigation Satellite Systems	306
11.2	Quasi-Zenith Satellite System	306
11.2.1	Overview	306
11.2.2	Constellation	307
11.2.3	Signals and Services	308
11.2.4	Spacecraft	313
11.2.5	Control Segment	317
11.2.6	Operations Concept	319
11.2.7	Current Performance	320
11.3	Indian Regional Navigation Satellite System (IRNSS/NavIC)	321
11.3.1	Constellation	322
11.3.2	Signal and Data Structure	322
11.3.3	Spacecraft	327
11.3.4	Ground Segment	330
11.3.5	System Performance	333
	References	334

following sections. The system architecture, service provision including navigation signal properties and service performance to be provided, as well as the deployment plan or schedule are mentioned for each system. Additionally, initial demonstration results are presented.

Applications of global navigation satellite systems (GNSS) are getting more widespread and deeply penetrated into our daily life and economy. Having their own national space-based positioning, navigation, and timing (PNT) infrastructure has thus become a natural interest of many countries. Independent PNT services are an important aspect of national security and sup-

port economical growth through sharing the growing market of GNSS applications. However, it is quite difficult for most countries to deploy a fully global satellite navigation system, since this requires a high level of technologies in a broad variety of fields as well as a huge financial budget.

11.1 Concept of Regional Navigation Satellite Systems

As a minimum, a satellite navigation system requires simultaneous observations of four satellites to obtain precise position, velocity, and timing at the user location. If the system covers the entire surface of the Earth, several dozens of satellites are indispensable. One of the solutions for realizing a system with minimum cost thus consists in limiting the service area, that is, to establish not a *Global*, but a *Regional* Satellite Navigation System (RNSS).

The use of geosynchronous orbits (including both geostationary orbits, GEOs, and inclined geosynchronous orbits, IGSOs, as well as high eccentricity Earth orbits (HEOs)), which have the same or orbital period as the Earth's rotation (i.e., one sidereal day) or half this value, is the most effective way to establish an RNSS at the minimum number of satellites. In the early stage of European satellite navigation system studies, that is, prior to the decision for a global medium Earth orbit (MEO) constellation, the combination of GEO and HEO satellites in Molniya or Tundra orbit was considered as a candidate constellation for a regional system, which could later be extended to a fully global system [11.1]. Molniya orbits, which have a half sidereal day orbital period, were designed to provide long duration of visibility for a high latitude area including polar regions [11.2]. The Molniya satellite system has been used in the former Soviet Union as a communication satellite system since mid-1960. The Tundra orbit is a derived orbit, which has a one sidereal day orbital period, approximately 23h 56min. Use of IGSO satellites has also been considered as an alternative to traditional GEO satellites for various types of communication and navigation services over the contiguous United States [11.3, 4].

In Japan, the Communications Research Laboratory (CRL), which is the precursor of the current National Institute of Information and Communications Technology (NICT), proposed *Figure-8 satellites* for mobile communications as well as satellite navigation in the late 1990s. The *Figure-8 satellite* uses

a circular IGSO and the footprint of the satellite draws a figure of eight on the Earth's surface [11.5]. Japanese industry made a proposal after the concept study for a Japanese Regional Navigation System (JRANS, [11.6]), and the Japan Aerospace Exploration Agency (JAXA) also investigated extension of the current Quasi-Zenith Satellite System (QZSS) to an RNSS in the early 2000s [11.7].

The performance of a satellite navigation system in terms of accuracy, availability, continuity, and integrity has a strong correlation with the number of satellites as well as the satellite geometry (as described by the so-called geometric dilution of precision, GDOP). The utilization of geosynchronous satellites is an effective choice for providing four or more satellites' visibility with a minimum number of satellites for a limited regional service area, since the satellites have good visibility and can be observed for a longer time period in the respective region.

Assuming that there are only four satellites in the sky, the best GDOP can be obtained under the condition that one satellite is located in the zenith and three others at -19.47° elevation separated by 120° in the azimuth plane [11.8]. Obviously, this condition cannot be realized in practice and the designers of an RNSS system are required to allocate satellites in the constellation such as to obtain an optimum GDOP at all times throughout a desired service area [11.9].

In this chapter, two RNSSs are introduced in the following sections. To obtain an ideal satellite geometry throughout Japan, QZSS is using IGSO satellites in order to always put a satellite near the zenith, while GEO satellites are planned to be allocated for lower elevations in the future expansion. IRNSS/NavIC, on the other hand, has a different approach to obtain good geometry over the Indian subcontinent, which is located near the equator. Here, GEO satellites are visible near the zenith, while IGSO satellites cover lower elevations and provide the desired spacing in the azimuth plane.

11.2 Quasi-Zenith Satellite System

11.2.1 Overview

The QZSS is a regional space-based PNT system, which has been deployed by the Japanese government since 2003 [11.10]. At the beginning of the development, the primary intention was to augment the US GPS in Japan by a three-satellite constellation using IGSOs.

The satellites in the constellation have the same orbital period as a satellite, that is, 23h 56min, but a higher orbital inclination and slight eccentricity. These orbital parameters were optimized for maintaining better visibility at high elevation angles in Japan. This is of particular interest for urban canyons and mountainous terrain, where navigation users cannot receive a suffi-

cient number of navigation signals from GPS satellites alone. Each of the three IGSO satellites in the constellation flies over Japan for a period of eight hours and at least one satellite provides navigation signals at more than 60° elevation.

The QZSS was named after these orbital characteristics. The first satellite of QZSS, known as QZS-1 or *Michibiki* (a nickname given in an open call by the public), was launched on September 11, 2010. The technical verification and application demonstrations were conducted successfully by the JAXA and other research institutes. Taking account of the successful results, the Japanese government announced on September 30, 2011 to establish a four-satellite QZSS constellation as a national infrastructure by the late 2010s, as well as to set the future goal of a seven-satellite constellation by around 2023 [11.11] to maintain an independent PNT capability at minimum cost.

The National Space Policy Secretariat (NSPS) within the Cabinet Office took responsibility for the deployment of the operational system. The procurement process for the establishment of the ground control segment and the service provision through 15 years from 2018 to 2033 as a Private Finance Initiative (PFI) project started in 2013. A group led by NEC Corporation was selected as service provider and QZSS Service Corporation (QSS) was established to implement the PFI project. As for the satellite procurement, Mitsubishi Electronic Corporation (MELCO) was awarded the manufacturing contract for three additional spacecraft.

The requirements for the operational QZSS define the following services [11.12, 13]:

- **GPS Complementary Service:** L1 C/A and L1C signals on L1 (1575.42 MHz), L2C signal on L2 (1227.6 MHz), and L5 signal on L5 (1176.45 MHz) will be transmitted. Those signals have the highest interoperability with GPS and other GNSSs by using the same radio frequency as well as the same message structure and format as GPS.
- **GPS Augmentation Service:** Two types of the augmentation service will be provided. One is the submeter level augmentation service (SLAS) for code-phase positioning users and the other is the Centimeter Level Augmentation Service (CLAS) for carrier-phase positioning users.
- **Public Regulated Service:** Authorized users can access the dedicated public signal transmission service, which uses encryption and provides users with more signal security. In case of GPS jamming or spoofing, the service can provide independent positioning and timing information.
- **Early Warning Service:** Using a special message type defined in the submeter level augmentation sig-

nal, short messages are to be transmitted for early warning in the case of natural disasters.

- **Message Communications Service:** A satellite communications link will be made available for safety confirmation among families and employees in companies just after natural disasters such as a huge earthquake. As part of the QZSS Safety Confirmation Service (Q-ANPI), the user-generated messages are relayed to the control center through a geostationary QZSS satellite and subsequently forwarded to their final destination by e-mail.

The above-mentioned services will start in the beginning of April 2018, following the launch of three additional QZSS satellites, which are scheduled for the 2017 time frame [11.11, 12]. At the time of writing (mid-2015), the system definition and total system design study are still ongoing. The following sections, therefore, describe the demonstration system for the first satellite *Michibiki* as a reference.

11.2.2 Constellation

The QZSS consists of a combination of IGSO and GEO satellites. For the basic four-satellite constellation, which is to be completed in 2018, three IGSOs and one GEO satellite will be adopted. All of these satellites have a 23^h56^m orbital period synchronized with the Earth's rotation period. Accordingly, the QZSS visibility conditions repeat after one sidereal day.

The three IGSO satellites are placed in three distinct orbital planes with a 120° offset in their respective right ascensions of the ascending node (RAAN). In addition, their arguments of latitude are also separated by 120°. This choice results in an identical groundtrack for all three satellites, while their equator crossing times differ by one-third of an orbital period. Accordingly, the satellites pass over the same region with an 8 h separation and offer a continuous 24/7 availability of one satellite at high elevation angles.

In order to optimize the visibility in the area surrounding Japan, the orbits have an inclination of $i = 43^\circ$ as well as a slight eccentricity ($e = 0.075$). Its apogee is placed in the northern-most point of the orbit, that is, the argument of perigee is $\omega = 270^\circ$. The RAAN and mean anomaly are selected such that the center longitude of the ground track is located at 135° East. Figure 11.1 shows the QZSS orbits in an inertial reference frame and its ground track relative to the surface of the Earth. As a result of the slight eccentricity and the specific choice of the perigee, the figure-of-eight ground track is not fully symmetric with respect to the equator. Due to the lower angular velocity near apogee, the satellites remain in the Northern Hemisphere for more than

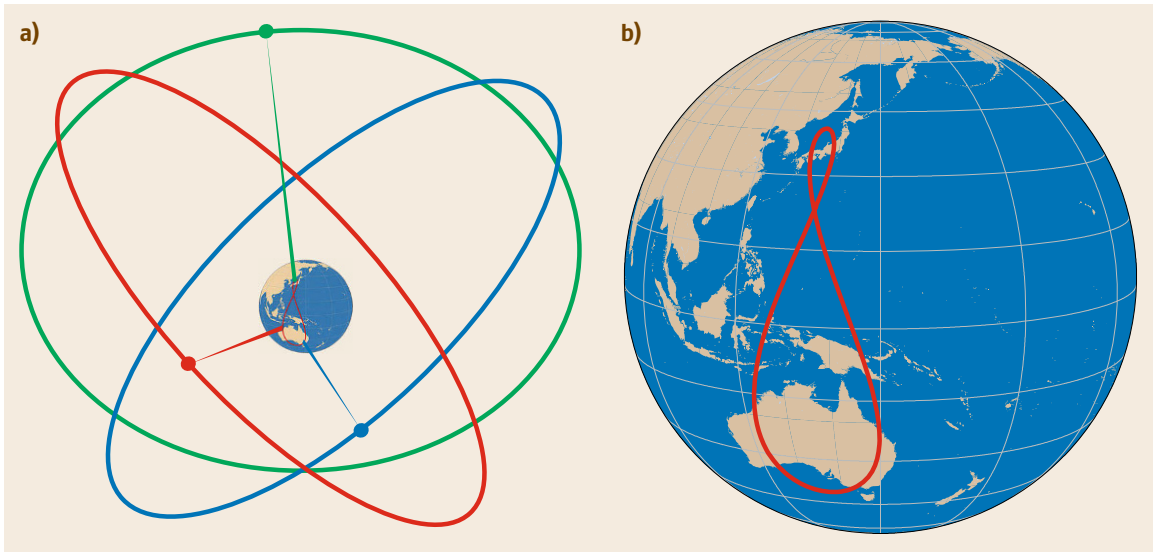


Fig. 11.1a,b QZSS IGSO constellation (a) (adapted from [11.14]) and ground track (b)

half of a day but quickly passes the perigee arc over Australia. As discussed in [11.15], this asymmetrical figure-8 orbit provides a good compromise in terms of service availability, link properties, and robustness between a fully symmetric orbit and a *tear-drop* orbit that have been considered as alternatives in the system design studies. The reference orbital elements for the QZSS IGSO constellation as initiated by *Michibiki* (QZS-1) are summarized in Table 11.1.

Due to Earth oblateness, the orbital plane of the QZSS IGSO satellites performs a slow precessional motion and the longitude of the ascending node decreases at an average rate of $3.65^\circ/\text{y}$. In addition, the orbits are subject to luni-solar perturbations, which induce long-term changes in the orientation of the orbital plane. In order to maintain the center longitude of the ground track within the desired deadband, small orbit correction maneuvers are performed approximately twice per year to adjust the mean motion of the QZSS satellites. Furthermore, out-of-plane maneuvers are required to compensate the secular changes of the orbital inclination [11.16].

Table 11.1 QZSS orbital parameters (reference values; [11.14])

Element	Value
Semimajor axis a	42 164 km (average)
Eccentricity e	0.075 ± 0.015
Orbital inclination i	$43^\circ \pm 4^\circ$
Argument of perigee ω	$270^\circ \pm 2^\circ$
Central long. of ground track $\bar{\lambda}$	$135^\circ \pm 5^\circ$ East
RAAN spacing $\Delta\Omega$	120°

Based on simulations of the long-term evolution of the QZS-1 orbit, the initial RAAN of this satellite has been chosen such as to minimize the inclination change over the expected lifetime. This has helped to notably reduce the required fuel for orbital corrections on the *Michibiki* spacecraft, but fixes the RAAN values as well as the total inclination change and fuel demand for the subsequent IGSO satellites.

11.2.3 Signals and Services

The QZSS satellites transmit a variety of navigation signals to support the applications and services introduced in Sect. 11.2.1. Four of these signals provide the basis of the GPS complementary service and are fully interoperable with legacy and modernized civil GPS signals in the L1 (1575.42 MHz), L2 (1227.60 MHz), and L5 (1176.45 MHz) frequency bands (Table 11.2). In addition, two augmentation signals for the SLAS and the CLAS are transmitted on L1 and E6 (1278.75 MHz), respectively, which provide error-correction messages for code-phase positioning and high-precision carrier-phase positioning. This subsection briefly summarizes the properties of these navigation signals. Further details of the QZS-1 signals are defined in the interface specification for QZSS users [11.14]. Signals and services launched in 2018 will be defined in a separate document [11.19–21].

Aside from the L-band navigation signals, QZSS GEO satellites will make use of S-band signals (near 2 GHz) for the Q-NAPI Safety Confirmation Service. This service will enable users to forward short safety

Table 11.2 Overview of QZSS navigation signals [11.14] and planned modifications for the operational services starting in 2018 [11.11, 17–21]

Signal	Channel	Band	Center frequency (MHz)	Minimum user received power (dBW)	Category	Modifications for operational service
L1-C/A		L1	1575.42	−158.5	GPS interoperable signal	
L1C	Data channel	L1	1575.42	−163.0	GPS interoperable signal	
	Pilot channel	L1	1575.42	−158.2	GPS interoperable signal	
L2C	Time multiplexed ^a	L2	1227.60	−160.0 (total)	GPS interoperable signal	
L5	I channel	L5	1176.45	−157.9	GPS interoperable signal	Additional L5S signal for experimental use on separate signal
	Q channel	L5	1176.45	−157.9	GPS interoperable signal	
L1-SAIF		L1	1575.42	−161.0	Augmentation signal	Followed by L1S signal which provides SLAS and message service, and L1Sb for multi-function satellite augmentation system (MSAS) follow-on service ^b
LEX	Time multiplexed ^a	E6	1278.75	−155.7 (total)	Augmentation signal	Followed by L6 signal which provides CLAS

Notes:

^a The L2C and LEX signals employ interleaved bit streams for concurrent transmission of two independent ranging sequences on a common physical channel.^b The L5Sb augmentation signal will be made available from QZS-2 onward. The L1Sb signal will be provided from a GEO satellite in the QZSS constellation.

information when other communication links are unavailable in the case of a natural disaster.

GPS Interoperable Signals

A primary motivation for the QZSS lies in the enhancement of the restricted GDOP of GPS-only navigation in urban-canyons and mountainous areas. The signals serving this *GPS Complementary Service* are designed to minimize any required modifications on the user equipment. Receivers shall be able to seamlessly acquire and track both GPS and QZSS signals, decode their navigation messages, and calculate user position, velocity, and time from the combination of GPS and QZSS observations.

In accord with the current and planned civil navigation signals of GPS, the QZSS transmit the legacy L1 C/A signal, the modernized civil L2C and L5 signals (supported by the GPS Block IIR-M and/or IIF satellites), as well as the L1C signal (to be available from GPS Block III onward). It may be noted that QZSS does not support the transmission of L1/L2 P(Y)-code signals, which are most widely used today for dual-frequency navigation by today's geodetic GPS receivers and serve as a basis for the clock offset determination of GPS satellites. This introduces minor conceptual differences in the handling of GPS and QZSS clock information and requires a proper consideration of intersignal biases by the user.

The technical parameters of these GPS-interoperable signals are intended to be the same as those of the corresponding GPS signals. This includes their radio-frequency (RF) properties as well as their navigation message structure and format as defined in the latest version of the GPS signal specifications (i.e., IS-GPS-200 [11.22] for L1C/A and L2C, IS-GPS-705 [11.23] for L5, and IS-GPS-800 [11.24] for L1C). In the case of the L1 C/A, L2C, and L5 signals, QZSS employs identical modulation schemes and code generators to GPS, but different pseudorandom noise (PRN) sequences. At present, PRN numbers 193–202 have been reserved by the GPS directorate for the transmission of GPS-like navigation signals by the QZSS constellation [11.25].

As a small difference between GPS and QZSS signal structures, QZS-1 uses a pure BOC(1,1) modulation [11.26] for the L1C signal instead of the modified binary offset carrier modulation (MBOC) that has been selected for GPS following the GPS/Galileo signal optimization efforts conducted by the United States and the European Union. Here, the pilot channel of the L1C signal is modulated with a BOC(1,1) subcarrier for 29 out of 33 code chips, whereas the remaining 4 of 33 chips are modulated with a BOC(6,1) waveform (Chap. 7 and [11.27]). In addition to the slightly different choice of the subcarrier, QZS-1 also uses a 90° phase shift between the L1C data and pilot channels [11.14], whereas both channels are phase aligned in GPS. For the new

Block II QZSS satellites a phase alignment fully consistent with GPS will be employed [11.19].

The joint transmission of three distinct L1 signal components (L1 C/A as well as L1C data and pilot) requires a special interplex modulation, which includes a so-called intermodulation (IM) product to achieve a constant signal power envelope irrespective of the individual signal and subcarrier states (Chap. 4). Following [11.28], the resulting signal at time t can be described as

$$S(t) = (A_{L1CA}S_{L1CA} + A_{L1CD}S_{L1CD}) \cos(\omega t) + (A_{L1CP}S_{L1CP} + A_{IM}S_{L1CP}S_{L1CA}S_{L1CD}) \sin(\omega t) \quad (11.1)$$

or

$$S(t) = A \sin \left(\omega t + S_{L1CP}S_{L1CA}\beta_1 + S_{L1CP}\frac{\pi}{2} + S_{L1CP}S_{L1CD}\beta_2 + \frac{3}{2}\pi \right) \quad (11.2)$$

with

$$\beta_1 = \tan^{-1} \left(\frac{A_{L1CA}}{A_{L1CP}} \right) \quad \text{and} \quad \beta_2 = \tan^{-1} \left(\frac{A_{L1CD}}{A_{L1CP}} \right). \quad (11.3)$$

Here, $S_i = c_i s_i d_i = \pm 1$ with $i = L1CP, L1CD$, and $L1CA$ denoting the product of ranging code c_i , subcarrier s_i (not on L1 C/A), and navigation data d_i (not on L1C pilot) for the L1C pilot, L1C data, and L1 C/A-code signals. Furthermore, A_i denotes the corresponding signal amplitudes and ω represents the L1 carrier frequency.

In accord with the relative signal powers given in Table 11.2, values of 44.2° and 30° have been adopted for the angles β_1 and β_2 in the QZS-1 interplex signal, respectively. They correspond to almost equal amplitudes of the C/A and L1C pilot component, and an L1 data channel amplitude that is roughly 40% smaller (and almost identical to the amplitude of the intermodulation product). The resulting in-phase/quadrature (IQ) constellation diagram is illustrated in Fig. 11.2.

The LNAV, CNAV, and CNAV2 navigation messages transmitted by QZSS as part of the L1 C/A, L2C/L5, and L1C signals, respectively, are designed for maximum communality with GPS. However, some unavoidable difference arise due to the different purpose and implementation of the QZSS. Among others, the different orbit of the QZSS satellites requires a different choice of reference values for the semimajor axis, eccentricity, inclination, and nodal drift in the almanac and ephemeris parameters. Also, some flags and

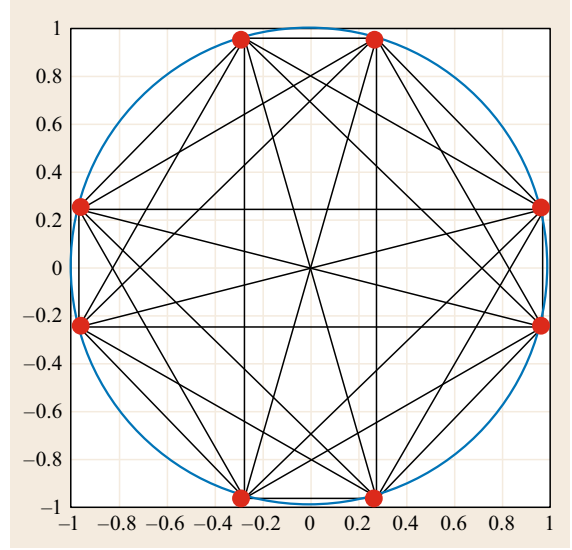


Fig. 11.2 IQ diagram of the QZS-1 interplex signal comprising L1 C/A as well as the L1C data and pilot channels. Image credits: F. Antreich

signal-related parameters (e.g., group delay parameters) require a different content or interpretation than that in GPS. Finally, the QZSS exploits the flexibility of the civil navigation message (CNAV) and CNAV2 concept, by defining various new messages that are not required in a standalone navigation system. As an example, QZS-1 can provide retransmitted almanac, ionosphere, and time-offset parameters for the GPS satellites in addition to the native QZSS data [11.14].

Due to the specific characteristics of their orbit, the QZSS IGSO satellites are in permanent access of the master control station (MCS) and can be provided with updated navigation data on a rapid time scale. New clock and ephemeris data of QZS-1 are, for example, issued once every 15 min as compared to 2 h for GPS. Along with highly stable onboard clocks, this contributes to a very favorable signal-in-space range error of QZSS (Sect. 11.2.7).

The complete compatibility and interoperability between QZSS and GPS have been confirmed by GPS-QZSS Technical Working Group jointly established by the US and Japanese governments [11.29]. Here, compatibility refers to the fact that the two navigation systems will not interfere with each other in a harmful manner, while interoperability addresses the ability to jointly receive and use the signals of both systems in a single receiver.

L1-SAIF Signal

Visibility restrictions in urban areas not only affect the tracking of GNSS satellites, but also likewise con-

strain the reception of ranging error corrections and integrity data from satellite-based augmentation systems (SBAS; Chap. 12). Traditionally, satellite-based augmentation system (SBAS) services are provided through geostationary satellites, which can host the SBAS transponder as a secondary payload and offer a continuous uplink capability from a fixed ground station. SBAS signals from geostationary satellites can well be received by aviation users with full-sky visibility, but are often unavailable for land mobile users due to their moderate elevation. A car running along an east–west street in an urban environment will commonly face severe signal interruptions, when buildings on the southern (or northern) side of the street obstruct the view to the SBAS satellite.

Due to its unique orbit, QZSS offers an ideal platform to relay differential GPS (DGPS) error-correction messages to mobile users in urban canyons. The continuous visibility of at least one high elevation satellite can overcome the aforementioned difficulties and users can obtain higher accuracy and reliability than normal standalone GPS navigation when using the augmentation signals.

The L1-SAIF (L1 submeter class augmentation with integrity function) signal design [11.14, 30] matches that of other SBAS signals and is based on a BPSK(1) modulation with a 1023-chip Gold-code ranging sequence of 1 ms duration. For the L1-SAIF signal, dedicated PRN numbers in the range of 183–192 have been reserved [11.25], which are different from the PRN used for the GPS-compatible ranging signals transmitted via the main L-band antenna. As an example, QZS-1 transmits standard navigation signals as PRN 193, but employs PRN 183 for the L1-SAIF signal.

The L1-SAIF signal is phase coherent with the L1 C/A and L1C signals but transmitted via a different antenna and amplifier to avoid an overly complex modulation scheme. The overall L1 spectrum and IQ constellation are presented in [11.31] based on signals analysis with a high-gain antenna.

In accord with the SBAS standard [11.32], the L1-SAIF employs a data message with a symbol rate of a 500 sps and a 1/2 forward error correction (FEC) yielding an effective data rate of 250 bps. Each message has a length of 250 bits including an 8 bit preamble, a 6 bit message identifier, 212 data bits, and a 24 bit cyclic redundancy check (CRC) field. For the provision of GPS range corrections as well as ionospheric delay and integrity information, the QZSS SAIF signal utilizes the same messages as traditional SBAS systems. This ensures good backward compatibility with existing SBAS and facilitates the build-up of L1-SAIF receivers from existing SBAS receivers.

On the other hand, QZSS-specific messages are required to provide QZSS orbit information in the SAIF data stream to support a self-contained ranging function independent of the GPS-type navigation signals. Due to the higher orbital inclination and eccentricity, the standard GEO ephemeris message (message type MT9) cannot be used with QZSS and is substituted by a new MT58 ephemeris message. Its concept is motivated by the GLONASS orbit model (Sect. 3.3.3) and the message provides a Cartesian state vector as well as perturbational accelerations. These data are then used to numerically integrate the QZSS trajectory from the given initial conditions. Aside from the QZSS ephemeris message, SAIF-specific messages are also defined for tropospheric corrections, intersignal biases, and selected other parameters. A comprehensive description of these messages is provided in the QZSS signal interface control document (ICD) [11.14].

Correction data transmitted in the L1-SAIF signal are generated in real-time based on observations of the Japanese GPS Earth Observation Network System (GEONET). This network of continuously operating reference stations (CORS) is operated by the Geospatial Information Authority of Japan (GSI) and comprises GNSS receivers in more than 1200 locations throughout Japan. For QZS-1, data from the GEONET monitoring stations are processed at the L1-SAIF Master Station (L1SMS) hosted by the Electronic Navigation Research Institute (ENRI) in Tokyo. The resulting SAIF messages are then forwarded to JAXA's MCS for uplink to the satellite [11.33, 34].

The SLAS offered by QZSS aims at providing DGPS correction messages to users throughout Japan. The SLAS correction data are designed for single-frequency GPS L1 C/A-code users and enable submeter positioning accuracy [11.35, 36], which represents a notable improvement over the GPS Standard Positioning Service (SPS) performance (typically 5 m three-dimensional (3-D) rms).

LEX Signal

Besides the L1, L2, and L5 frequencies utilized by GPS, QZSS also transmits navigation signals in the Galileo E6-band centered at 1278.75 MHz. As indicated in Table 11.2, these signals serve the future Public Regulated Service (PRS) and the centimeter level augmentation service (CLAS). Both services will gradually be implemented along with the build-up of the QZSS IGSO and GEO constellation, but an *L-band Experimental* (LEX) signal was already transmitted on QZS-1 to conduct preparatory experiments and technology demonstrations for the CLAS. The main characteristic of this signal is its high data transmission rate of 2 kbps,

which is realized through code shift keying (CSK) modulation.

The LEX (or E6b) baseband signal is generated with a chipping rate of 5.115 MChip/s. Its spectrum matches that of a BPSK(5) signal with a 5 Mcps PRN code, but is actually generated by interleaving two 2.5575 Mcps bit streams on a chip-by-chip basis [11.14, 37]. These comprise

- A 4 ms PRN short code with a length of 10 230 chips that is modulated by means of CSK with the Reed–Solomon (RS)-encoded navigation message, and
- A 410 ms PRN long code with a length of 1 048 575 chips that is modulated by square wave with a period of 820 ms beginning from 0 (010101...).

Similar to the GPS/QZSS L2C signal, a data channel and a pilot channel are thus combined in the LEX signal using time multiplexing. An overview of the signal generation is provided in Fig. 11.3.

The PRN codes for both channels are Kasami sequences [11.38], which are generated from the combination of a 20 bit and a 10 bit linear feedback shift register (LFSR) [11.14]. While the 10 bit register is common to both code generators, distinct 20 bit registers (albeit with identical feedback taps) are used for the two PRN codes. While the short code is truncated at 10 230 chips, the length of the long code corresponds to the maximum length sequence of a 20 bit LFSR. However, the long code is reset once at the start of a week, since the 820 ms repeat rate is not commensurable with the length of a day or week.

The LEX navigation message transmitted via the data channel has a total length of 2000 bits, and one complete message is transmitted each second. The LEX message is composed of a 49 bit header (including the 32 bit preamble as well as the PRN number (8 bits), the

message number (8 bits), and a 1 bit alert flag), a 1695 bit data part and 256 parity bits. The latter are based on Reed–Solomon (RS) encoding of the data bits and the 17 header bits that follow the preamble [11.14]. With the given number of parity bits, errors in up to 16 8-bit data-plus-header symbols can be corrected before an unrecoverable frame error occurs. Following [11.39], a zero frame error rate was achieved above 10° elevation during LEX trials conducted in the Tokyo area. For a different receiver, using the L1 C/A-code signal as a reference for the CSK demodulation, a 90% probability of proper frame decoding is reported in [11.37, 40] for observations above 40° elevation based on LEX trials conducted in Melbourne, Australia.

The data rate (2 kbps) of the LEX navigation message is substantially higher than that of traditional direct-sequence spread spectrum (DSSS) signals used for GNSS navigation. At the 4 ms code duration, such signals could at best support a data rate of 250 bps. The exceptional data rate is achieved through the use of CSK modulation, which effectively treats the data component of the LEX signal as a communication channel [11.41].

The concept of CSK modulation is illustrated in Fig. 11.4. Considering a navigation message made up of a series of 8 bit symbols, the CSK modulation creates a shifted copy of the native PRN sequence, where the byte value N_i of the i th data symbol specifies the number of chips, by which the PRN code sequence is shifted during the given data symbol. In the case of QZSS, the duration of a data symbol matches the 4 ms code period. The CSK modulation thus allows the transmission of 8 bits during one code period rather than a single bit in traditional DSSS signals.

While the CSK modulation is clearly favorable in terms of the data rate, it necessitates the availability of an independent pilot channel for synchronization, since

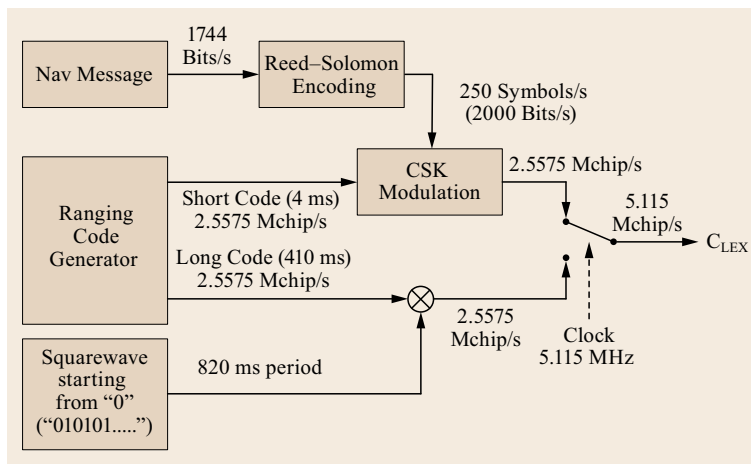


Fig. 11.3 QZSS LEX signal generation [11.14]

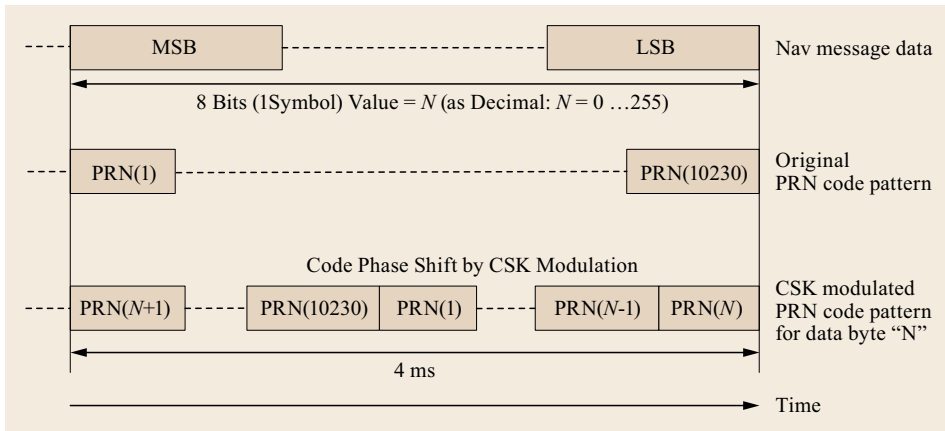


Fig. 11.4 CSK modulation of the LEX data channel [11.14]

the cyclically shifted PRN code of the data channel would otherwise be very difficult to acquire [11.41]. Also, the CSK demodulation is more demanding in terms of receiver architecture than that of conventional DSSS signals and only a very limited number of receivers presently (2015) support a data extraction from the LEX signal.

The high data rate of the LEX signal provides the basis for the dissemination of real-time correction data for carrier-phase-based precise point positioning (PPP; Chap. 25). During the QZS-1 demonstration phase, different augmentation methods were tested by several Japanese organizations. Both GSI and the Satellite Positioning Research and Applications Center (SPAC) conducted experiments involving the transmission of carrier-phase correction data via the QZSS LEX signal. The GSI experiment aimed to facilitate surveying in urban areas with single-frequency network-based RTK [11.42], while SPAC targeted dual-frequency applications in surveying, precise farming, and construction machinery control [11.43].

For dual-frequency precise point positioning (PPP) applications, the LEX signals support the transmission of correction data in the State Space Representation (SSR) format defined by the Radio Technical Commission for Maritime Services (RTCM) [11.44]. RTCM-SSR messages comprise orbit and clock corrections for GNSS satellites relative to the respective broadcast ephemerides. Other than range corrections employed by wide-area augmentation systems, the SSR corrections are not restricted to regional users but can be applied on a global scale [11.45]. They have emerged as a new standard for real-time PPP users and can be provided through various communication channels. As an example, RTCM-SSR correction data for GPS and GLONASS are provided by the International GNSS Service (IGS) to its users [11.46] using real-time internet streaming. QZSS offers the unique possibility to

transmit such corrections (with a suitably reduced volume) along with the navigation signal [11.47]. These facilitate real-time PPP in remote areas that are not well covered by other communication links. As part of the LEX experiments, JAXA also offers SSR corrections for multiconstellation PPP users [11.37, 48].

The experiments and technology demonstrations conducted with the QZS-1 LEX signal will provide the basis for CLAS of the fully operational QZSS. As described in [11.20], the new Block II QZSS satellites will provide two data channels using CSK modulation, thus offering a higher capacity for correction data.

11.2.4 Spacecraft

As of 2015, the QZSS is made up of a single demonstration satellite which transmits most of the planned navigation signals and helps to validate diverse augmentation concepts. Even though QZS-1 will ultimately become a full member of the final QZSS constellation, the follow-on satellites are independently developed and will differ in various aspects from the QZS-1 design. Nevertheless, the QZS-1 satellite is well suited to illustrate the basic elements of the spacecraft platform and relevant payload elements.

Platform

The first QZSS satellite (QZS-1, or *Michibiki*; Fig. 11.5) is designed based on JAXA's Engineering Test Satellite-VIII (ETS-8, [11.49]) and the derived DS2000 GEO bus system of Mitsubishi Electric Corporation (MELCO). By building on an established spacecraft design, the QZS-1 development cycle could be notably reduced and the use of a GEO spacecraft platform offered ample resources for the accommodation of the complex and power-intensive navigation payload.

QZS-1 was launched on September 11, 2010 with an HII-A rocket from Tanegashima Space Center and



Fig. 11.5 The first satellite of QZSS *Michibiki*. The artist’s drawing illustrates the location of the main L-band navigation antenna (L-ANT) and the complementary SAIF antenna (LS-ANT) as well as the Ku-band antenna for bi-directional time comparison and the telemetry, tracking and commanding (TT&C) antenna (courtesy of JAXA)

injected into an elliptical transfer orbit. It was then raised into its final 33 000 km × 39 000 km orbit through multiple firings of its apogee boost motor.

Key characteristics of the spacecraft are summarized in Table 11.3. Compared to other GNSSs using MEOs, the QZS-1 satellite requires higher power and a large amount of propellant for injection into the geosynchronous orbit. Overall, QZS-1 carried about 2.3 t of fuel, which amounts to more than half the total mass at lift-off.

In order to ensure a sufficient life-time covering both the initial demonstration phase and the subsequent use as part of the fully operational QZSS constellation, QZS-1 has been designed with due redundancy. The spacecraft has dual independent electrical bus systems, high-capacity lithium ion batteries, and excess solar panel size. These features allow us to continue minimum operations with GPS interoperable signals in the case of a single electrical bus failure. Furthermore, QZS-1 offers a specific back-up configuration of the attitude control subsystem to continue operations in the case of attitude sensor failures [11.50].

The attitude and orbit control system (AOCS) of QZS-1 employs redundant star sensors, Sun sensors, and Earth horizon sensors for attitude determination. An orbit propagator fed with uploaded orbit information is used to translate between the inertial and orbital frames and to steer the rotation angle of the solar panels [11.50]. Four reaction wheels (complemented by thrusters for wheel unloading) serve as actuators for orientation changes [11.51].

Similar to other GNSS satellites, the QZSS satellite needs to point the navigation antenna toward the Earth throughout its orbit, while orienting the solar pan-

Table 11.3 QZS-1 satellite system key parameters

Parameter	Value
Dry mass	1800 kg (total satellite system) 330 kg (nav. payload)
Wet mass	4100 kg (at separation)
Dimensions	2.9 m × 3.1 m × 6.2 m (Body) 25.3 m (Span)
Design life time	>10 years
Solar array power	5300 W @ 10 years
Pointing Accuracy	0.1°
TT&C	5000–5010 MHz (uplink) 5010–5030 MHz (downlink)
Reliability	> 0.8 (Bus) > 0.7 (Payload)

els to the Sun. Depending on the Sun elevation above the orbital plane (the so-called β -angle), either the *yaw-steering* attitude control modes or the *orbit-normal* mode is employed by the QZS-1 satellite to achieve this goal (Fig. 11.6).

In yaw-steering mode, which is also most widely used by GNSS satellites in MEOs (such as GPS and GLONASS), a continuous rotation about the Earth-pointing (yaw) axis is performed to keep the solar panel rotation axis perpendicular to the Sun and Earth direction. In this way, the surface normal of the solar panels can always be aligned with the Sun direction, which maximizes the effective cross-section and thus the received energy. As a disadvantage, however, large and rapid yaw-slews near local noon and midnight are required during periods of low Sun elevation (Sect. 3.4 of this Handbook), which may exceed the capabilities of the reaction wheels. QZS-1, therefore, adopts the

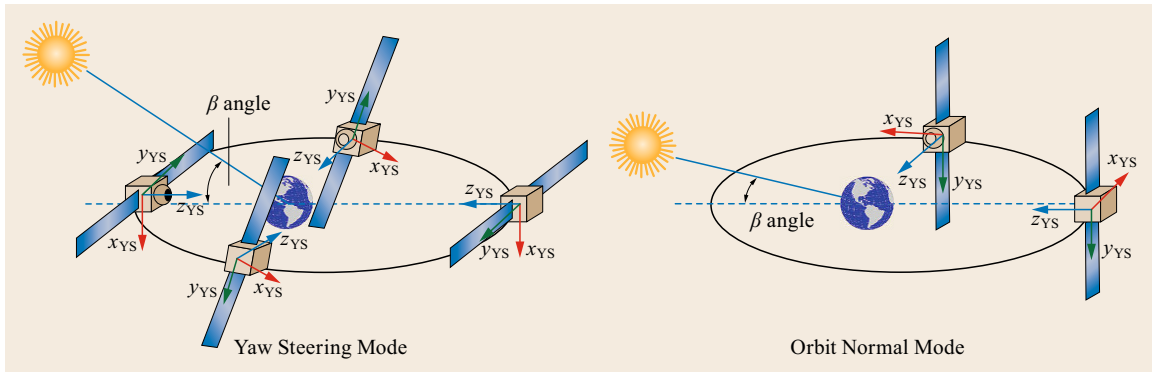


Fig. 11.6 Attitude control modes of the QZS-1 spacecraft (adapted from [11.51])

orbit-normal mode for $|\beta| < 20^\circ$. Here, the solar panel rotation axis is maintained perpendicular to the orbital plane. This avoids the need for yaw rotations, but yields a slightly reduced power output due to the suboptimal pointing of the solar panels.

Details of the QZS-1 attitude control system and the implementation of the two attitude modes are described in [11.51]. Mathematical models describing the nominal attitude of the QZS-1 spacecraft as a function of its orbital position and the Sun direction are given in [11.52]. As pointed out in [11.53], care must be taken, though, that the transition between yaw-steering and orbit-normal mode does not take place at $|\beta| = 20^\circ$ exactly. Instead, the mode switch is conducted near this threshold at an orbital position that minimizes the required yaw-angle change during the mode transition.

Payload

In accord with the overall QZSS mission goal, the navigation payload [11.54] constitutes the primary payload element of QZS-1. However, considering that QZS-1 is the first satellite in the IGSO, it also has been equipped with various environmental sensors as a secondary payload. The Technical Data Acquisition (TEDA) experiments serve to better characterize the IGSO environment and, if needed, to adapt the design of follow-on QZSS satellites [11.50]. More specifically, the TEDA package comprises a light particle telescope (including the alpha particle and proton sensor-B, APS-B, and the electron sensor-A, ELS-A) as well as magnetometers (MAM) and potential monitors (MAM). A detailed description of these sensors and the initial flight results is given in [11.55].

The navigation payload of QZS-1 comprises three individual subsystems:

- The *L-band signal transmission subsystem* (LTS) is the main part of the navigation payload, which generates the reference clock and baseband signal as

well as the navigation message, and then modulates, amplifies, and transmits the navigation signals.

- The *Time transfer subsystem* (TTS) developed by the NICT enables two-way satellite time and frequency transfer (TWSTFT), comparison of the onboard atomic clock with a ground-based frequency standard, and a remote synchronization of an onboard crystal oscillator (RESSOX) from the ground [11.56, 57]. The measurements and control commands are exchanged through a dedicated bi-directional Ku-band communication link.
- The *laser reflector assembly* (LRA) is a fully passive device that enables high-accuracy distance measurements by satellite laser ranging (SLR).

Figure 11.7 provides a block diagram of the navigation payload on QZS-1. The onboard reference clock is generated by the time-keeping system, which comprises the Rubidium Atomic Frequency Standard (RAFS), a time-keeping unit, a synthesizer, and a navigation onboard computer.

The RAFS selected for QZS-1 has been manufactured by PerkinElmer and is identical to the rubidium clocks of the GPS IIF satellites [11.58] that are known for their superior stability. For redundancy purposes, QZS-1 is equipped with two independent RAFS units. Even though a space-capable hydrogen maser for use within the QZSS program has been developed by NICT [11.59], the use of rubidium clocks was ultimately preferred in view of their lower mass and power consumption [11.56].

The time-keeping unit also includes a voltage- and oven-controlled crystal oscillator (VCOCXO) with high short-term stability. This is steered by the navigation computer to follow the RAFS frequency on average. In this way, a clock signal with high stability over a wide range of time scales is achieved. As an alternative, the VCOCXO can also be controlled from the ground through the TTS. This has been demon-

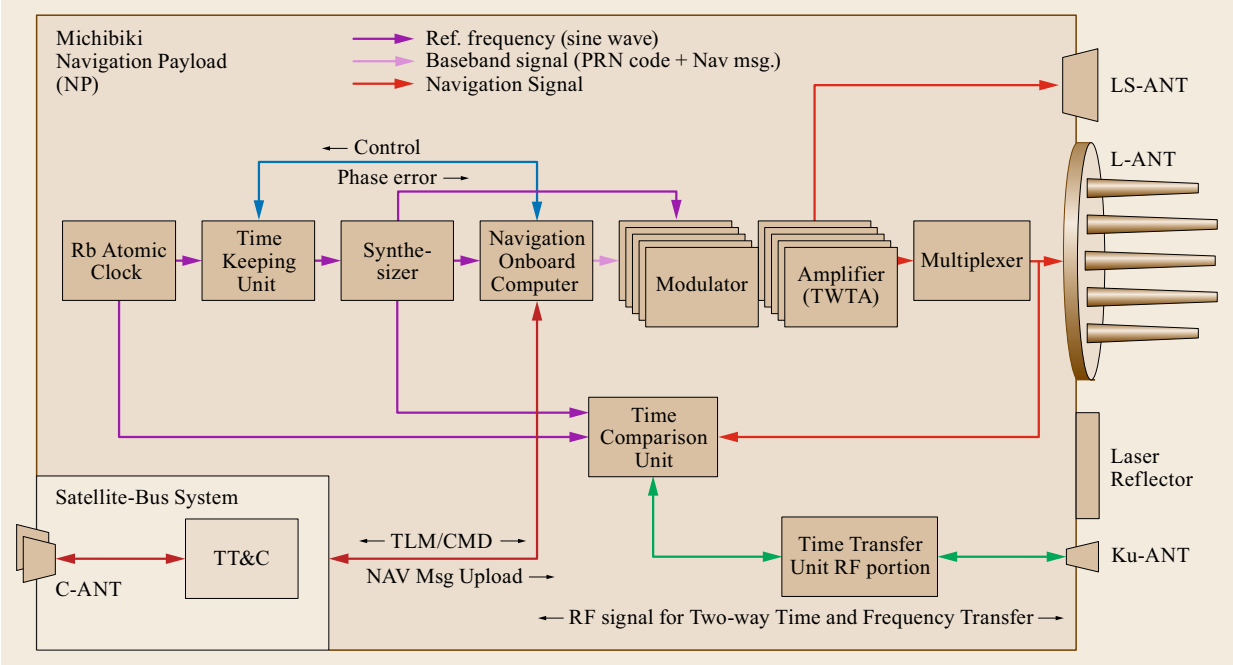


Fig. 11.7 Block diagram of QZS-1 navigation payload

strated with the Remote Synchronization System of the Onboard Crystal Oscillator (RESSOX) described in [11.60]. Here, the QZS-1 reference oscillator is steered to follow a highly stable ground clock with subnanoseconds accuracy. While the resulting short-term stability achieved in the RESSOX experiments is lower than that of the standard onboard clock, an excellent long-term stability of $4.4 \cdot 10^{-14}$ was achieved at time scales of 100 000 s.

Based on the clock reference of the time-keeping unit, the synthesizer finally outputs the satellite clock and L-band carriers. The navigation onboard computer then generates the baseband signals with the navigation message, which are subsequently amplified, multiplexed, and transmitted.

Since the orbit of QZS-1 is higher than that of traditional MEO constellations, QZS-1 needs a higher transmission RF power. Values for the signals transmitted via the main L-band antenna are given in Table 11.4. The individual signals are amplified through an adjustable channel amplifier and a traveling-wave tube amplifier before being combined at the multiplexer

(MUX). The peak power amounts to nearly 1 kW. In order to maintain enough margin with respect to high-power durability and thermal resistance, the L1-SAIF signal is transmitted through a different path and antenna than the other L1-band signals.

The main L-band antenna shown in Fig. 11.8 is composed of 19 individual helix antennas arranged in an outer ring (12 elements), an inner ring (6) elements, and a central element. The individual elements are phase coherently combined to form a gain pattern, which provides uniform signal strength on the Earth's surface. Compared to the boresight direction, the antenna gain is about 0.2 dB higher at 5° boresight angle

Table 11.4 QZS-1 radio frequency output power

Signals	Power (W)
L1 C/A & L1C	91
L2C	20
L5	70
LEX	63

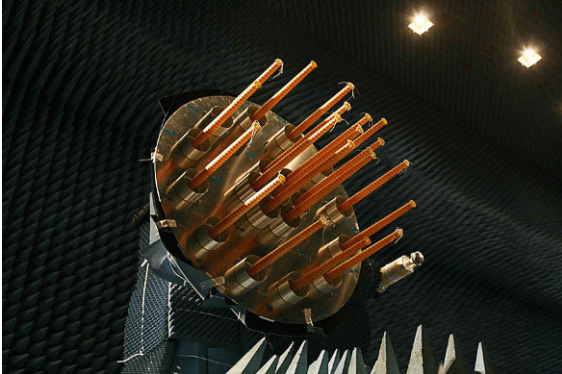


Fig. 11.8 QZS-1 L-band antenna array during anechoic chamber testing (courtesy of JAXA)

and 0.5 dB lower at 8.5° (corresponding to the Earth rim) for L1 signals [11.59].

Figure 11.9 illustrates the location of the various antennas and the laser retroreflector array on the Earth-facing panel of the QZS-1 satellite. The main L-band antenna is placed such that the phase center is located on the z -axis of the spacecraft body coordinate system passing through the spacecraft center of mass. In this way, the phase-center location is unaffected by the spacecraft attitude and only a radial phase-center offset needs to be considered in the GNSS data analysis.

The laser retroreflector is accommodated opposite to the SAIF antenna. It consists of a planar array of 7×8 uncoated corner cube prisms of 4 cm diameter [11.61]. Laser ranging observations are routinely conducted by observatories in Japan (Tanegashima, Koganei, Tokyo, Japan) and Australia (Yarragadee, Mt. Stromlo) as well as various other stations of the International Laser Ranging Service (ILRS; [11.62]). Even though the SLR observations of QZS-1 are primarily used for the validation of GNSS-based precise orbit determination results [11.63], they can also serve as independent means for standalone orbit determination [11.64].

11.2.5 Control Segment

The control segment for the first QZSS satellite is composed of an MCS, the monitoring stations (MS) network, the tracking control station (TCS), and the

tracking, telemetry, and command (TT&C) station. These are complemented by a time management station (TMS) for time measurement synchronization with external time scales.

This section provides descriptions of each subsystem of the control segment as well as the operation scenario and the current performance. Figure 11.10 shows the distribution of the control segment for the demonstration system.

Master Control Station

The MCS acts as the focal point of navigation-related QZS-1 operations. In particular, the MCS performs the following functions:

- Orbit and clock offset estimation and generation of navigation messages for upload to the satellite
- Monitoring and controlling of the navigation payload onboard the spacecraft
- Remote monitoring and controlling of the monitoring stations
- Monitoring of the navigation signal quality and user range error and generating health/alert flags in the case of malfunctions
- Handling of navigation messages and control commands generated by other research institutes as part of specific technology demonstrations and experiments.

The MCS is located at the Tsukuba Space Center (TKSC), approximately 60 km northeast of Tokyo downtown. Even though there is only a single MCS facility for the QZSS demonstration phase, it satisfies high-availability requirements through internal use of a hot-redundant system.

Monitoring Stations

Nine monitoring stations located in Japan as well as Asian and Oceania areas have been installed to receive the QZSS and GPS signals for precise orbit and clock estimation. A total of five monitoring stations outside Japan were established in collaboration with local organizations. These include National Aeronautics and Space Administration (NASA) (Kokee Park Geophysical Observatory, Kauai), National Ocean and Atmospheric Administration (NOAA) (Weather Forecast Office, Guam), Geoscience Australia (Mt. Stromlo), the Indian Space Research Organisation (ISRO) (ISRO Telemetry Tracking and Command Network, ISTRAC, Bangalore), and the Asian Institute of Technology (Geoinformatics Center, Bangkok).

The monitoring stations receive L-band navigation signals from the QZSS and GPS constellation and transmit raw data to the MCS via ground or satellite communication link. A QZSS receiver and a GPS

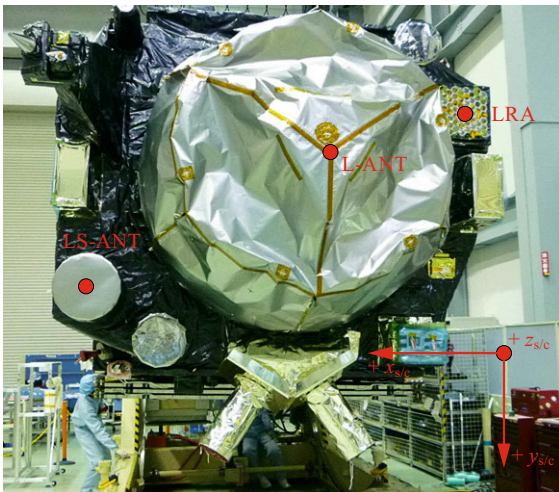


Fig. 11.9 Earth panel configuration of QZS-1 showing the main L-band antenna (L-ANT), the L1-SAIF antenna (LS-ANT), and the laser retroreflectors assembly (LRA). The 19-elements helical array antenna is covered with a thermal insulator in order to maintain proper thermal conditions on the orbit. Arrows indicate the orientation of the spacecraft coordinate system adopted by JAXA (courtesy of JAXA)

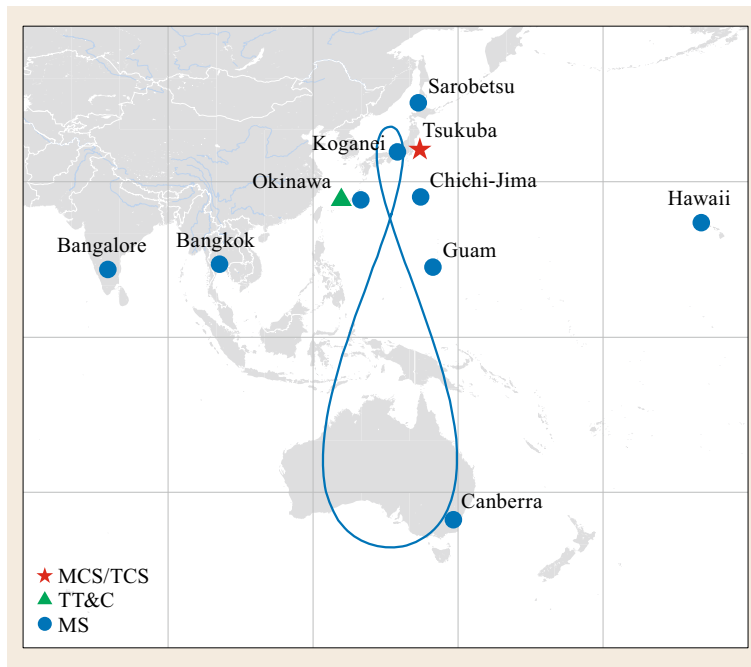


Fig. 11.10 Control segment distribution for the QZSS demonstration. For reference, the ground track of QZS-1 is indicated by a solid line

receiver with cesium atomic frequency standard and a multiband antenna with radome as well as meteorological sensors, local computer, and communication equipment are installed in each MS. It is noted that the use of independent receivers for QZSS and GPS signals implies the need for estimating an inter-receiver (inter-system) bias for precise orbit determination of QZSS and GPS in the MCS.

Tracking and Control Station

The TCS, which is collocated with the MCS at JAXA's TKSC, is in charge of the QZS-1 satellite bus operations. Among others, this covers the following functions:

- Performance of housekeeping operations, that is, monitoring of the spacecraft telemetry and issuing of commands for routine operation
- Remote monitoring and control of the TT&C station
- Planning and execution of orbit-keeping maneuvers and reaction wheel unloading
- Integration of all system operations as well as time and resource allocation between navigation payload and satellite bus system operation.

Due to the continuous visibility of the QZSS satellites, the GEO satellite operation system previously developed by JAXA could be applied to QZSS with minimum modifications.

From the early stage of the development phase, a model-based operation approach was adopted on the

operations system and procedure design. Routine operations are fully automated so that less operators are required to control the satellite in an efficient manner [11.65, 66].

Tracking, Telemetry, and Command Station

The main TT&C ground station for QZS-1 is located in the Okinawa main island, where QZSS is permanently visible and continuous operations can be ensured. Since the Okinawa district is regularly exposed to typhoons during the summer season, two redundant high-gain antennas with a diameter of 7.6 m and a protecting radome (Fig. 11.11) are installed at the TT&C station to ensure uninterrupted operation even in severe weather condi-



Fig. 11.11 Tracking, Telemetry, and Command Station (TT&CS) in Okinawa island (courtesy of JAXA)

tions. The TT&C station is operated remotely from the TCS in Tsukuba.

Time Management Station

The TMS [11.57, 67] measures the difference between the QZS-1 onboard clock and a highly stable ground reference and conducts TWSTFT operations. Based on the TMS measurements, the coordinated universal time (UTC) offset parameters and the GPS-QZSS Time Offset (GQTO) for the QZS-1 navigation message are generated.

The TMS is operated by NICT, the Japanese authority, in charge of generating the national standard time. For redundancy purposes, the TMS comprises independent facilities at Koganei and Okinawa, each of which comprises a Ku-band antenna, an atomic clock, and a time-comparison equipment.

The TMS at Koganei is connected with UTC(NICT) and colocated with one of the QZSS monitoring stations. The receiver clock of the Koganei monitoring station is defined as QZSS reference time (QZSST) and the TMS measures the offset between this time scale and UTC(NICT). In addition, the Koganei TMS performs time and frequency comparisons with the United States Naval Observatory (USNO) in Washington via a TWSTFT link, which involves two GEO communication satellites and a relay station at Kogee Park, Kauai. The offset between UTC(NICT) and UTC(USNO) as measured by TWSTFT was used for the evaluation of the GQTO.

A secondary TMS is deployed at NICT's Subtropical Environment Remote-Sensing Center in Okinawa. While not directly connected to UTC(NICT), this TMS offers a larger antenna (3.7 m diameter as compared to 1.8 m in Koganei) and continuous 24 h visibility of QZS-1 [11.56]. Synchronization of both TMS facilities is achieved through TWSTFT.

11.2.6 Operations Concept

As a key function, the MCS performs a real-time estimation of the orbits and clock offsets for all active GPS and QZSS satellites. The process makes use of a square-root information filter and processes pseudorange and carrier-phase measurements of the (presently nine) monitoring stations at 30 s intervals [11.68]. Aside from the GNSS orbit and clock parameters, the filter state comprises the clock offset values of the monitoring stations as well as carrier-phase biases and station-specific tropospheric delays. Furthermore, empirical solar radiation pressure parameters are estimated in view of the limited accuracy of available a priori models. Overall, a total of about 500 parameters is adjusted in each measurement update.

Following [11.68], the orbit determination process is designed to recover the actual orbit with an accuracy of better than 0.7 m in the radial direction and twice this value in along-track and cross-track directions. Slightly larger values of 0.8 and 1.6 m apply for the associated orbit prediction, accuracy over a 2.5 h interval. For clock offset determination and 35 min prediction accuracy requirements of 2.5 and 4.4 ns have been specified. The targeted limit for the total signal in space range error amounts to 1.6 m, but is well outperformed in practice.

QZSS orbit information is referred to a realization of the Japan Geodetic System (JGS, [11.69]) maintained by JAXA based on GNSS observations of 40 reference stations (including the 9 QZS-1 monitoring stations as well as stations of the IGS and complementary SLR observations). The latest frame realization is known as JGS2010 and designed to be rigorously aligned with ITRF2008.

Clock offsets are initially determined relative to the QZSS system time scale (QZSST), which is defined by the receiver in the Koganei monitoring station connected with Japanese reference time standard, UTC(NICT). However, the time offset between QZSST and GPS time (GPST) is included into the clock offset parameters transmitted in the navigation message, so that user equipment can calculate positions without explicitly considering the time difference between both systems.

For the ionospheric error correction of single-frequency users, the same Klobuchar model as in GPS is used [11.14, 70]. However, the model coefficients for the QZSS NAV message are generated such as to best match the vertical total electron in the area surrounding Japan rather than aiming at a global fit. Observations from 300 nationwide CORS stations are used to adjust the Klobuchar parameters and the coefficients for the navigation message are frequently updated to better reflect the dynamic behavior of the ionosphere.

Nominally, the orbital ephemeris and clock parameters are updated every 15 min, while ionospheric correction parameters are updated once every hour. In contrast to this, the augmentation messages with state space or ranging error corrections, integrity, and other value-added information are generated and transmitted via QZSS in (near) real time. During the QZS-1 demonstration phase, several types of augmentation methods were tested and evaluated as described in Sect. 11.2.3.

As a regional navigation satellite system with a small number of satellites, care must be taken to minimize PNT service interruptions due to orbit maintenance maneuvers and thruster activity for reaction wheel unloading. The interval between two orbital maintenance maneuver is required to be more than 150

days, and wheel unloadings shall be separated by at least 40 days.

Orbit corrections for satellites in IGSO are required on a routine basis due to a secular increase of the semimajor axis induced by the Earth's triaxiality. For QZS-1, the perturbation amounts to roughly 85 m/d and results in a westward acceleration of the mean sub-satellite longitude. To ensure that the center longitude remains in the specified range of $135^{\circ}\text{E} \pm 5^{\circ}$, periodic orbit maneuvers are conducted, in which the along track velocity and thus the semimajor axis are decreased. The orbit correction maneuver imposes an initial ground track drift in the Eastern direction, which is gradually stopped and reversed by the natural orbital perturbations within several months. The velocity decrement is calculated based on long-term orbital dynamics simulations so that the center longitude of satellite ground track would not go beyond the specified range at the next orbit maintenance maneuver. Overall, the orbit maintenance strategy for the QZSS IGSO satellites closely matches the east–west station keeping of geostationary satellites [11.71] but employs less frequent maneuver due to the much wider control window.

For QZS-1, orbit corrections are performed roughly once every six months. In order to correct both the semimajor axis and the eccentricity, each correction is split into three individual maneuvers with a total magnitude of 1–4 m/s which are performed within one orbital revolution. For semimajor axis control alone, a velocity change of about $\Delta v \approx 0.5$ m/s is required.

Concerning angular momentum build-up in the reaction wheels, the first four years of QZS-1 operations have shown a similar profile in each year. Based on the collected experience, the time interval between two momentum wheel unloading operations could be extended to up to 150 days. Various momentum wheel unloadings could thus be combined with orbit maintenance operations.

In the case of any maintenance operations, users are informed about the expected outages and the renewed service availability through *Notice Advisory to QZSS Users* (NAQU) messages. For QZS-1, these notifications are made available online through JAXA's *QZ-Vision* homepage [11.72]. Planned maneuver information is also incorporated into the real-time orbit and clock determination process in the MCS. The filter can thus converge to its nominal performance in less than 21 h after orbit-keeping maneuvers and 9 h after a standalone wheel unloading [11.68].

11.2.7 Current Performance

Michibiki was launched from Tanegashima Space Center at 11:17 UTC) on September 11, 2010. In the

subsequent commissioning phase, the function and initial performance of each subsystem and the total system were checked out [11.73, 74], and the precise orbit determination (POD) software in the control segment was tuned. The L1 C/A and L2C signals were set healthy on June 22, 2011, followed by the L1C and L5 signals on July 14, 2011. All interface specifications and performances defined in the IS-QZSS were confirmed to be satisfied. In this section, the routine performance of the broadcast ephemeris and clock stability is described.

SIS-URE

The signal-in-space use range error (SIS-URE) describes the error in the modeled pseudorange caused by errors in the broadcast orbit and clock information. It is monitored in real time within the QZSS MCS. Each second, the instantaneous SIS-URE value is calculated from an average of the observation residuals at the nine QZS-1-monitoring stations using an elevation-dependent weighting.

Monthly performance reports are made available through JAXA's *QZ-vision* website [11.75]. By way of example, Fig. 11.12 shows the SIS-URE variation for July 2014. The root-mean-square (RMS) value in this period amounts to 0.34 m, which is well within the specified 95 percentile performance limit of ± 2.6 m. Compared to GPS, the QZSS clearly benefits from the continuous observations, the permanent upload capability, and the short (15 min) update interval of the broadcast navigation messages.

An independent assessment of the QZSS L1 C/A legacy navigation message (LNAV) broadcast ephemerides has been performed in [11.76] based on comparison with post-processed orbit and clock products of JAXA. Here, monthly RMS SIS-URE values

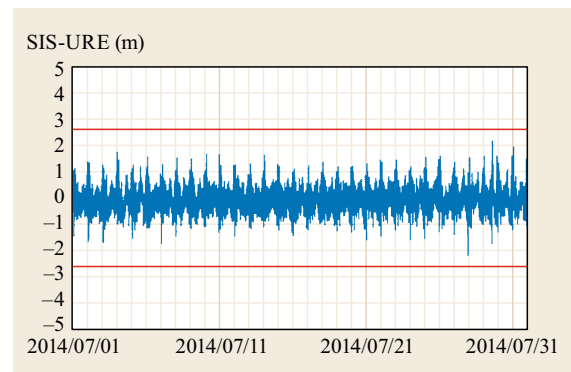


Fig. 11.12 QZS-1 SIS-URE as derived from observations of the QZSS real-time monitoring station network for July 2014. The red line marks the 95 percentile threshold value of ± 2.6 m specified for the QZS-1 routine operations (excluding orbit maintenance periods)

of 0.6 ± 0.2 m have been obtained for a one year analysis period starting March 2013. While these results indicate a slightly inferior performance than the MCS monitoring, they represent average errors over the entire globe rather than errors in the vicinity of the monitoring QZS-1 network.

Clock Stability

As discussed in Sect. 11.2.4, QZS-1 is equipped with a TTS, which has been employed in various experiments of the NICT to assess and the stability of the QZS-1 onboard frequency standards [11.57, 67].

On the other hand, the performance of the RAFS is evaluated on routine basis using L-band observations of the QZSS MS-monitoring stations. For the clock stability assessment over short time scales (typically less than 1000 s) the one-way carrier-phase (OWCP) method [11.77] is applied. Here, the Allan deviation (ADEV) is derived from detrended carrier-phase measurements collected at a monitoring station that is connected to a highly stable reference clock. For QZS-1, observations from the Okinawa site have been used, which offers the most stable frequency standard within the QZSS-monitoring station network. The results shown in Fig. 11.13 demonstrate a stability of $3 \cdot 10^{-12}$ at 1 s averaging intervals and about $5 \cdot 10^{-14}$ at 1000 s. Compared to [11.78], superior results at very short time are achieved due to the reduced impact of receiver noise in single-frequency observations.

For longer averaging times, QZS-1 clock offset solutions obtained as part of the precise orbit and clock determination have been employed. These are routinely generated by the MCS within six days based on the observation of the QZS-1-monitoring station network. At

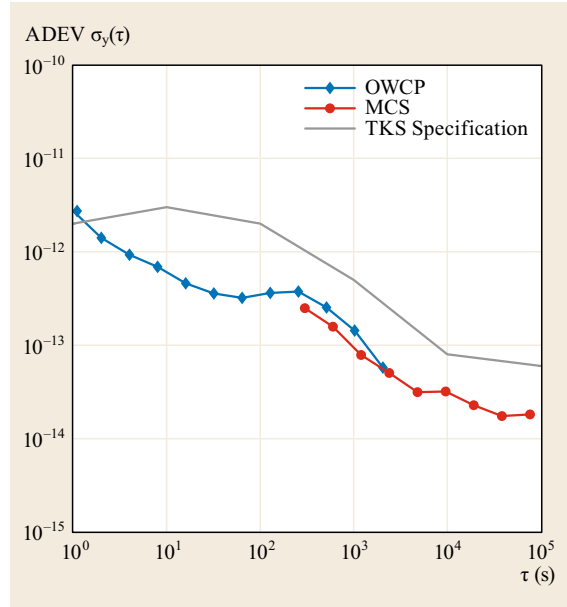


Fig. 11.13 Clock stability (Allan deviation, ADEV) of QZS-1 as determined with the one-way carrier-phase (OWCP) method and the precise orbit and clock determination by the QZSS master control station (MCS) in Aug. 2014. For comparison, the gray line indicates the specified performance of the time keeping system.

a one day time scale, a stability of $1 \cdot 10^{-14}$ is achieved (Fig. 11.13) in good accord with the inflight performance of the GPS Block IIF RAFS. Despite the small bump in the ADEV near 400 s, which is attributed to the time-keeping system, the overall performance of the QZS-1 clock stability is well within the specification and all time scales.

11.3 Indian Regional Navigation Satellite System (IRNSS/NavIC)

The IRNSS is an initiative by the ISRO to establish and operate an independent satellite-based navigation system, which provides services to the users in the region covering India and an area extending up to 1500 km from its geo-political boundary [11.79]. IRNSS is also known by the operational name NavIC (an acronym for Navigation with Indian Constellation) which means sailor/navigator in Sanskrit. This section on IRNSS describes its architecture, the various segments, and the signal and data structures. It provides a comprehensive picture of IRNSS as one of the emerging systems in the world of satellite-based navigation.

The IRNSS marks India's entry into the realm of independent satellite-based navigation systems. The ISRO is the key agency for the realization, operation,

and maintenance of the system. This involves building, launching, and operating the navigation satellites, as well as establishing the ground support systems.

The IRNSS is established with the objective of offering PNT services to the users in its service area. The system is designed to provide its users with a position accuracy of better than 20 m (2σ) in its primary service area.

The IRNSS classifies its service areas broadly into two regions, as shown in Fig. 11.14. The primary service area of IRNSS encompasses the Indian landmass and a region lying within a distance of 1500 km from its geo-political boundary. The secondary service area extends between latitudes 30°S to 50°N and longitudes 30°E to 130°E .

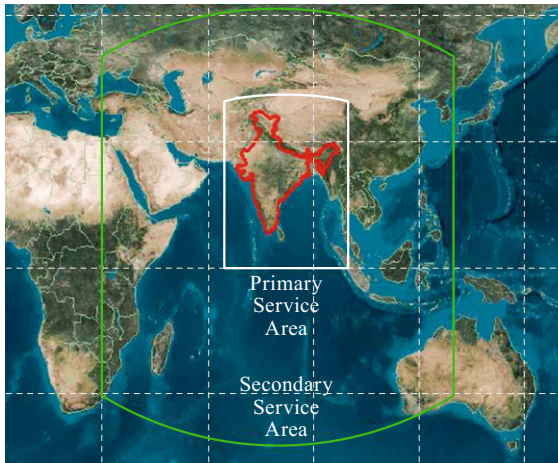


Fig. 11.14 IRNSS primary and secondary service areas (courtesy of ISRO)

IRNSS provides two types of navigation services:

- The *Standard Positioning Service (SPS)* is an unencrypted service provided to all the users within the IRNSS service area.
- The *Restricted Service (RS)* is an encrypted service provided only to the authorized users in the service area.

Similar to the other global and regional systems, the IRNSS navigation infrastructure can be divided into three segments illustrated in Fig. 11.15:

- The *Space Segment* consisting of the seven space vehicles that broadcast the IRNSS navigation signals to its users
- The *Ground Segment* consisting of the ground infrastructure to support the operation of the IRNSS system, such as the precise timing system, the spacecraft ranging mechanisms, the navigation software, the communication networks, and the spacecraft telemetry, tracking, and command network
- The *User Segment* consisting of the civilian and authorized users of IRNSS employing various types of IRNSS receivers.

Key components of the space and ground segment are described in the following subsections.

11.3.1 Constellation

The IRNSS space segment is made up of seven satellites, including three satellites in the GEO and four satellites in the IGSO with orbital planes tilted by 29° with respect to the equator. This constellation was selected based on the following design considerations:

- Minimizing the dilution of precision (DOP)
- Visibility of the maximum number of satellites over the targeted area
- Minimum satellite constellation
- System sustenance even in the case of a one-satellite failure, and
- Availability of orbital locations.

Given the large number of satellites (about 18 or up) required to ensure a guaranteed visibility of at least four satellites in an MEO constellation, the seven-satellite GEO and IGSO constellation was found to be optimum in view of the regional service requirements. In addition, the permanent 24 h visibility of all satellites over the region of interest facilitates the ranging, tracking, and commanding.

A summary of all satellites in the constellation is given in Table 11.5. The three GEO satellites are roughly equally spaced on the equator, covering a longitude range of about 100° between central Africa and Indonesia. The IGSO satellites, in contrast, describe figure-of-eight ground tracks covering a latitude range of about $\pm 30^\circ$ and centered at 55.0° and 111.75° East longitude, respectively (Fig. 11.16). The phasing of the satellites in the inclined orbits is chosen such as to avoid singular DOP values that might occur, if all IGSO s/c were to cross the equator at the same time [11.9].

11.3.2 Signal and Data Structure

The IRNSS utilizes the L5- and S-band frequencies allocated for Radio Navigation Satellite Services (RNSS). The carrier frequencies and the bandwidths of transmission are shown in Table 11.6.

The traditional navigation services frequency bands, that is, the L1- and L2-bands, have been completely utilized by the existing GNSS service providers. Hence, it was difficult to accommodate IRNSS in the L1- and L2-

Table 11.5 Satellites of the IRNSS constellation [11.80]

Satellite	Long.	Inclin.	Launched
IRNSS-1A	55.0°	$29^\circ \pm 2^\circ$	1 July 2013
IRNSS-1B	55.0°	$29^\circ \pm 2^\circ$	4 April 2014
IRNSS-1C	83.0°	$< 5^\circ$	15 October 2014
IRNSS-1D	111.75°	$29^\circ \pm 2^\circ$	28 March 2015
IRNSS-1E	111.75°	$29^\circ \pm 2^\circ$	20 January 2016
IRNSS-1F	32.5°	$< 5^\circ$	10 March 2016
IRNSS-1G	129.5°	$< 5^\circ$	28 April 2016

Table 11.6 Carrier frequencies and bandwidths

Signal	Carrier frequency (MHz)	Bandwidth (MHz)
L5	1176.450	24.0 (1164.45–1188.45)
S	2492.028	16.5 (2483.50–2500.00)

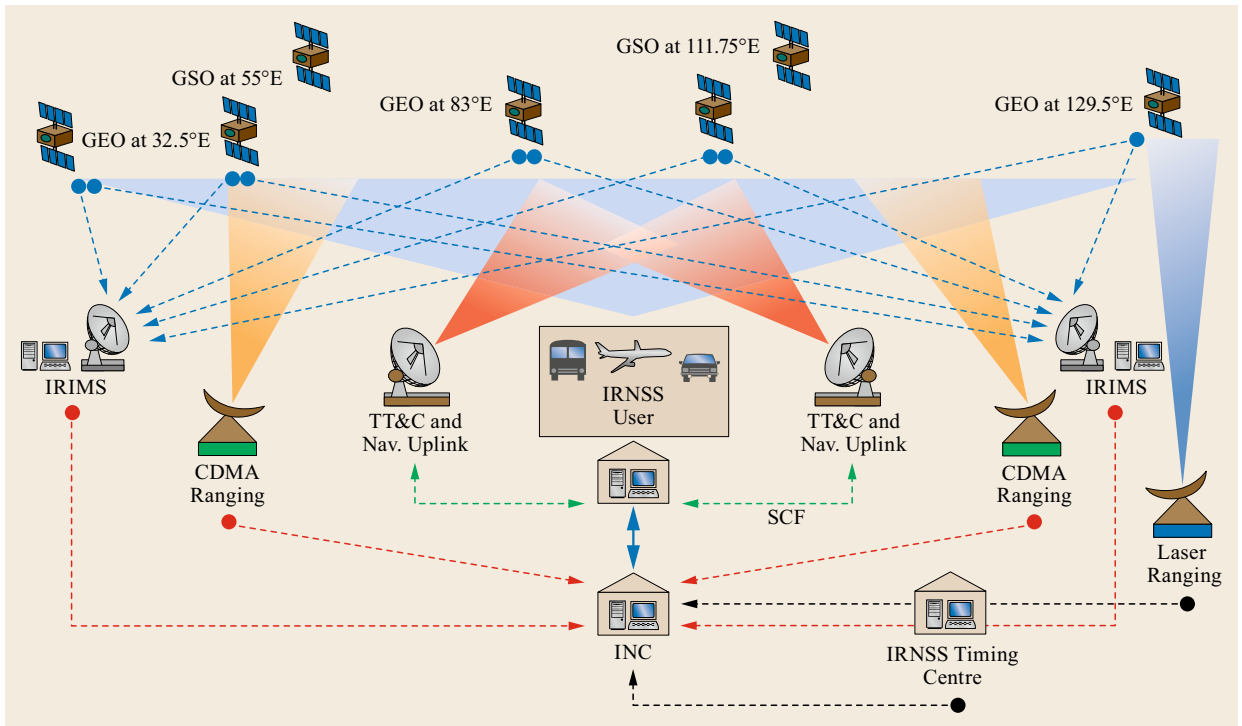


Fig. 11.15 IRNSS/NavIC architecture

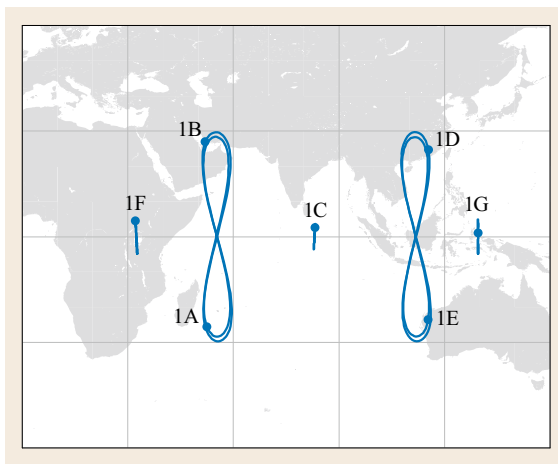


Fig. 11.16 IRNSS/NavIC constellation. The figure shows the ground tracks of the seven IRNSS satellites over the Indian Ocean region on 2 Aug, 2016 with circles indicating the position at midnight.

bands. The L5-band was chosen as it was less populated and as it was possible for IRNSS signals to co-exist with other GNSS signal in this band. This choice shall also facilitate interoperability with other GNSS signals using a common RF front end in the user receiver.

A second frequency was required to achieve better positioning accuracy using dual-frequency user receivers. The S-band has been allocated globally for radio determination satellite services (RDSSs) in 2012. Since function-wise there is very little difference between RDSS and RNSS, the International Telecommunications Union (ITU) has allotted a maximum bandwidth of 16.5 MHz for RDSS/RNSS. Hence, the S-band was chosen as it is a newly introduced band for navigation services, which was not yet populated. Also, S-band signals are less affected by ionospheric perturbations than the L-band.

Power levels of the IRNSS signals received on ground are chosen such as to co-exist with other GNSS signals operating in the respective frequency bands. They have been selected taking into account two primary considerations:

- The maximum power level of IRNSS shall not affect the reception of other GNSS signals and
- The minimum power level of the IRNSS signal shall provide sufficient signal strength to be detected even in the presence of other GNSS signals.

The resulting minimum and maximum power levels of IRNSS signals are given in Table 11.7.

Modulation

Each of the two carriers is modulated by three signals [11.81, 82]:

- The SPS data channel using a BPSK(1) binary phase shift key modulation with a chipping rate of 1.023 MHz
- The RS data channel using a BOC(5,2) binary offset carrier modulation with a 2.046 MHz ranging code and a 5.115 MHz subcarrier, and
- The RS pilot channel, which likewise uses a BOC(5,2) modulation.

When passed through a power amplifier or traveling-wave tube amplifier (TWTA) operated at saturation, the combination of these signals would produce a nonconstant envelope. Hence, an interplex product (Chap. 4) is added as a fourth signal component to achieve a constant envelope at the TWTA output.

The mathematical description for the three base-band navigation signals (i. e., the SPS binary phase shift keying (BPSK) data signal s_{sps} , the RS binary offset carrier (BOC) pilot signal s_{rs_p} , and the RS BOC data signal s_{rs_d}) is given by

$$\begin{aligned} s_{\text{sps}} &= \sum_{i=-\infty}^{+\infty} c_{\text{sps}}(|i|_{L_{\text{sps}}}) d_{\text{sps}}([i]_{D_{\text{sps}}}) \\ &\quad \times \text{rect}_{T_{c,\text{sps}}}(t - iT_{c,\text{sps}}), \\ s_{\text{rs}_p} &= \sum_{i=-\infty}^{+\infty} c_{\text{rs}_p}(|i|_{L_{\text{rs}_p}}) \\ &\quad \times \text{rect}_{T_{c,\text{rs}_p}}(t - iT_{c,\text{rs}_p}) sc_{\text{rs}_p}(t, 0), \\ s_{\text{rs}_d} &= \sum_{i=-\infty}^{+\infty} c_{\text{rs}_d}(|i|_{L_{\text{rs}_d}}) d_{\text{rs}_d}([i]_{D_{\text{rs}_d}}) \\ &\quad \times \text{rect}_{T_{c,\text{rs}_d}}(t - iT_{c,\text{rs}_d}) sc_{\text{rs}_d}(t, 0) \end{aligned} \quad (11.4)$$

with time t and the following definitions:

- $c_s(n)$ n th spreading code chip of signal s ,
- $d_s(n)$ n th navigation message chip,
- $sc_s(t)$ binary subcarrier,
- D_s number of chips per navigation data bit,
- L_s spreading code length in chips,
- $T_{c,s}$ spreading code chip duration.

The operations $|i|_x$ (i modulo x) and $[i]_x$ (integer part of i/x) provide the code chip index and the data bit

Table 11.7 Received power levels

Signal component	Maximum received power (dBW)	Minimum received power (dBW)
L5 SPS	−154.0	−159.0
S SPS	−157.3	−162.3

index for a given signal, while $\text{rect}_x(t)$ describes a rectangular pulse function of duration x . For the RS BOC signal, the subcarrier is defined as

$$sc_x(t, \varphi) = \text{sign}[\sin(2\pi f_{sc,x}t)], \quad (11.5)$$

where $f_{sc,x}$ denotes the subcarrier frequency. Individual signal parameters are summarized in Table 11.8. The choice of a BOC(5,2) modulation for the RS signal offers good spectral separation within the available bandwidth and is further discussed in [11.83].

A block diagram depicting the interplex signal generation is given in Fig. 11.17. Following [11.84] and [11.82], the composite signal at the carrier frequency f can be described by

$$\begin{aligned} s(t) &= \frac{\sqrt{2}}{3} [s_{\text{sps}}(t) + s_{\text{rs}_p}(t)] \cos(2\pi ft) \\ &\quad + \frac{1}{3} [2s_{\text{rs}_d}(t) - I(t)] \sin(2\pi ft), \end{aligned} \quad (11.6)$$

where the interplex signal

$$I(t) = s_{\text{sps}}(t) s_{\text{rs}_p}(t) s_{\text{rs}_d}(t) \quad (11.7)$$

is constructed in such a way as to realize the desired constant envelope.

Table 11.8 Parameter values for the IRNSS composite signal

Parameter	Unit	Value	Description
$R_{d,\text{sps}}$	sps	50	SPS data rate
$R_{c,\text{sps}}$	Mcps	1.023	SPS code chip rate
$R_{d,\text{rs}}$	sps	50	RS data rate
$R_{c,\text{rs}}$	Mcps	2.046	RS code chip rate
R_{sc}	Mcps	5.115	Subcarrier frequency

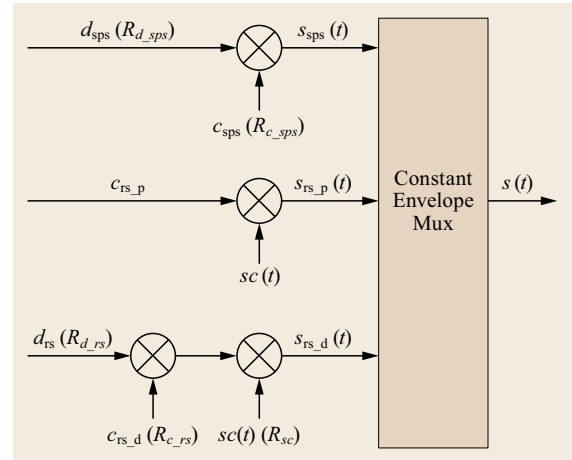


Fig. 11.17 Composite signal generation (after [11.84])

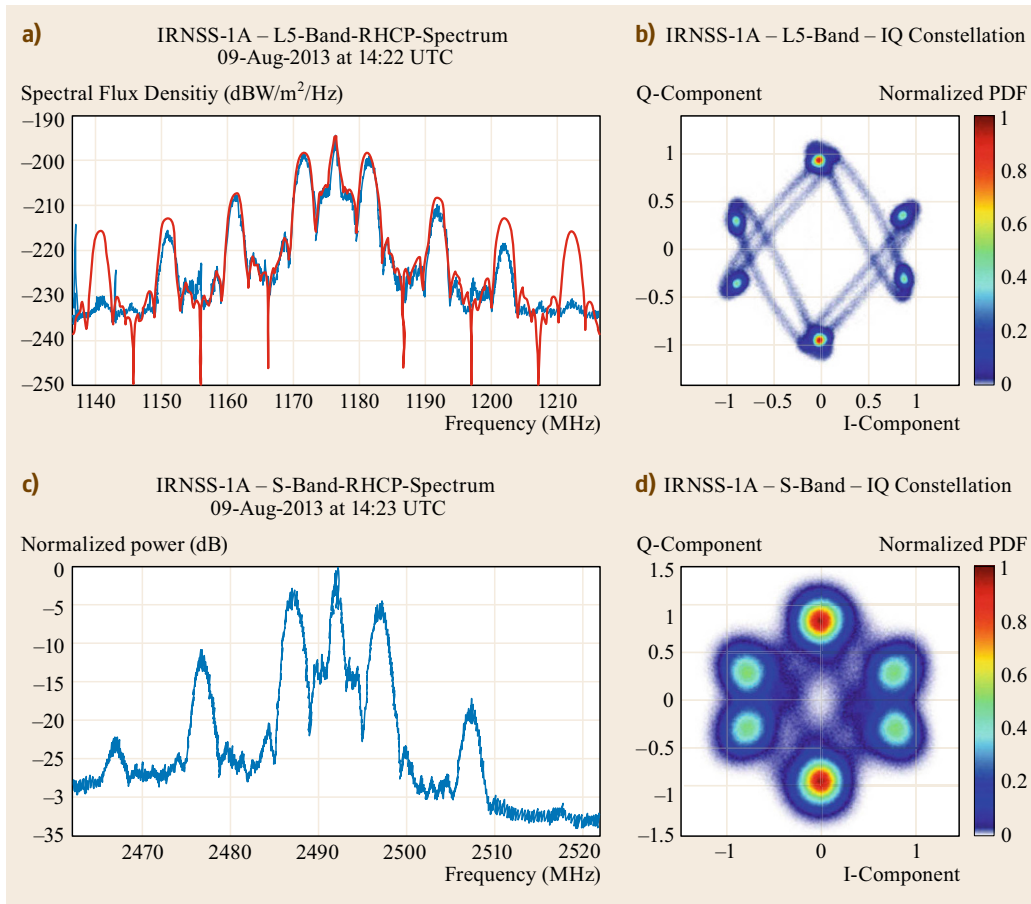


Fig. 11.18 IRNSS spectra (a,c) and IQ signal constellation (b,d) as obtained from the observations of the IRNSS-1A satellite with a high-gain antenna at the signal monitoring facility of the German Aerospace Center (DLR) on August 9, 2013 (after [11.82])

For the purpose of illustration, the spectra and IQ constellation of the L5- and S-band signals of the first IRNSS satellite are shown in Fig. 11.18. While the narrow peak at the center frequency of the spectra relates to the SPS BPSK(1) signal, the BOC(5,2) modulation generates two distinct lobes of 2 MHz width, which are separated by ± 5 MHz.

PRN Codes

PRN codes selected for the SPS are similar to the GPS C/A Gold codes [11.85]. The length of each code is 1023 chips and the code is chipped at 1.023 Mcps. Each spacecraft is allocated a unique PRN code and distinct codes are employed for the L5- and S-band, respectively.

The individual PRN codes are derived from two 10 bit maximum-length linear feedback shift registers (MLFSR) G1 and G2, which generate individual maximum length sequences (Fig. 11.19). In accord with GPS [11.22], the G1 and G2 generator polynomials are

defined as

$$G1: x^{10} + x^3 + 1,$$

$$G2: x^{10} + x^9 + x^8 + x^6 + x^3 + x^2 + 1.$$

Other than in GPS, however, which uses configurable taps to generate individual PRN code sequences, the IRNSS PRN codes are generated through different initial states of the G2 shift register to define the desired G2 chip delay. The resulting PRN sequence has a length of 1023 chips and is obtained from modulo-2 addition (xor-combination) of the G1 and G2 outputs. At the given chipping rate, the SPS code sequence repeats every 1 ms.

Initial values for the L5- and S-band PRN codes of the SPS and their association with the various IRNSS satellites are provided in Table 11.9.

For the regulated services, ranging codes with a length of 8192 chips and 4 ms duration are employed

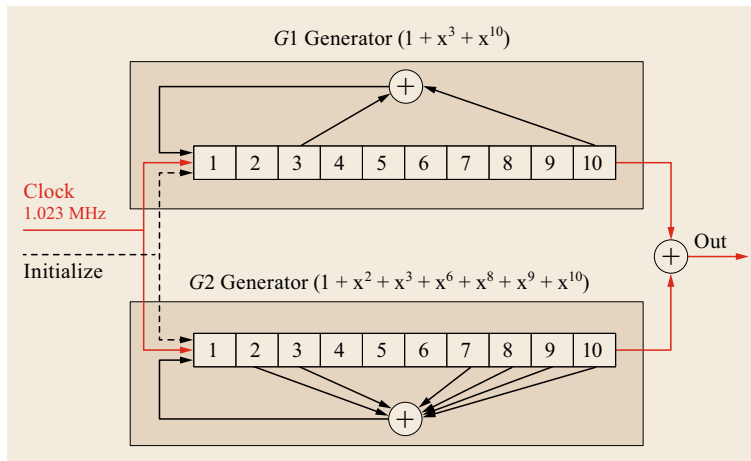


Fig. 11.19 IRNSS SPS code generator

as discussed in [11.86] and [11.82]. The pilot signal, furthermore, employs a secondary code of 40 chips, yielding a repeat period of 160 ms.

Navigation Data

The master frame of the IRNSS SPS navigation data comprise four subframes of 600 symbols length, which are transmitted at 50 sps [11.84, 86]. Each subframe has a 16 bit sync word followed by 584 symbols of forward error corrected and interleaved data. A 1/2 convolution encoding is applied for FEC of the navigation data, which results in a net amount of 292 bits per subframe (excluding the sync word). To protect against burst errors, the 584 symbols of the FEC-encoded navigation data are, furthermore, interleaved using a block interleaver with 73 columns and 8 rows. Data are first written in columns and read in rows. The start of each subframe is marked by a TLM word of 8 bits and each subframe ends with a 24 bit CRC to verify the data integrity of the received subframe bits as well as six additional tails bits.

The transmission time of each subframe is provided in the form of a time-of-week count (TOWC) following the telemetry (TLM) word. It is measured relative to the start-of-week and complemented by a week count transmitted in data of the first subframe. Both values refer to IRNSS System Time, which started at 23:59:47 UTC on August 21, 1999 and exhibits a constant nominal offset of 1024 weeks from GPS time.

The IRNSS navigation subframes employ a hybrid structure. Subframes 1 and 2 transmit a set of primary navigation parameters in a fixed structure. Subframes 3 and 4, in contrast, transmit secondary navigation parameters in the form of messages with varying structure and contents. As can be recognized from Table 11.10, the subframes 1 and 2 include a 232 bit data field, whereas subframes 3 and 4 employ a complementary 6 bit mes-

sage identification number and a slightly shorter data field.

The primary navigation data in subframes 1 and 2 include information, which is essential for the computation of a position fix and, therefore, continuously transmitted with a 48 s repeat rate:

- Satellite orbital elements
- Satellite clock correction model parameters
- Satellite and signal health status
- User range accuracy
- Total group delay.

They are complemented by the secondary navigation parameters transmitted in alternating messages with varying update rates in subframes 3 and 4:

- Satellite almanac
- Atmospheric (ionospheric) correction model
- IRNSS time offsets w.r.t. UTC and GNSS
- Constellation status
- Ionospheric grid delays and confidence
- Text messages
- Differential corrections
- Earth orientation parameters
- AutoNav-related parameters.

The navigation data are generated in the IRNSS Navigation Centre (INC) based on one-way and two-way ranging measurements, timing information from the IRNSS Network Timing Center (IRNWT), telemetry parameters, meteorological data, and ionospheric information. The clock and ephemeris data are estimated and then propagated for subsequent days. The primary navigation parameters are predicted for the subsequent 24 h and uplinked. IRNSS uses the world geodetic system (WGS) 84 coordinate system for position computation. The AutoNav parameters are clock

Table 11.9 Code assignment for SPS signals as defined in the IRNSS Signal ICD [11.84]. For each PRN, the satellite location, the initial value of the G2 Register, and the first 10 chips (in octal notation) are provided

PRN	Location	L5-SPS		S-SPS	
		initial G2	Chips	initial G2	Chips
1	55°E	1110100111	0130	0011101111	1420
2	55°E	0000100110	1731	0101111101	1202
3	83°E	1000110100	0713	1000110001	0716
4	111.75°E	0101110010	1215	0010101011	1524
5	111.75°E	1110110000	0117	1010010001	0556
6	32.5°E	0001101011	1624	0100101100	1323
7	129.5°E	0000010100	1753	0010001110	1561

Table 11.10 Structure of subframes 1 and 2 and 3 and 4

Bit index	Parameter	No. of bits
1	TLM	8
9	TOWC	17
26	Alert flag	1
27	AutoNav flag	1
28	Subframe ID	2
30	Spare	1
31	SF 1/2: Data	232
	SF 3/4: Message ID	6
	Data	226
263	CRC	24
287	Tail	6

and ephemeris data (similar to primary navigation parameters) predicted for seven days and uplinked to the spacecraft. In case the system is unable to uplink the primary parameters due to any reason, the AutoNav data are picked up from on-board memory and transmitted.

In the absence of data for a specific subframe, an idle pattern made up of alternating zeroes and ones is transmitted in the corresponding data field. In the case of missing primary navigation data in subframes 1 and 2, the transmission of the idle pattern is, furthermore, complemented by setting the alert flag in the subframe header [11.84].

Descriptions of the IRNSS navigation parameters and their application in the user receiver are provided in [11.87] as well as the IRNSS SPS ICD [11.84].

Based on various trade-off studies, a set of Keplerian elements as used in GPS and other MEO constellations has been adopted to describe the satellite orbits in the IRNSS broadcast ephemeris [11.87]. It may be noted that the same parameterization and orbit model are applied for both the IGSO and GEO satellites, whereas a special GEO model has, for example, been adopted for the regional BeiDou system to avoid potential singularities in the ephemeris generation at near-zero inclination [11.88].

The satellite clock offset from IRNSS System Time (also known as IRNWT) is described through a second-

order polynomial and an eccentricity-dependent correction for periodic relativistic effects. The resulting clock offset applies for dual-frequency L5/S-band users, and a *total group delay* parameter (T_{GD}) provided in the primary navigation parameters needs to be taken into account by single-frequency L5 users. Complementary intersignal corrections accounting for group delays between SPS and RS signals are discussed in [11.81]. Finally, translation from IRNWT to UTC and UTC(NPLI) (the UTC realization of the National Physical Laboratory of India) as well as the GNSS time scales of GPS, GLONASS, and Galileo is accomplished through dedicated messages transmitted at 20 min intervals.

Ionospheric information for single-frequency users is provided in the form of eight coefficients (α_i and β_i , $i = 1, \dots, 4$) for a Klobuchar-style single-layer model [11.89] as well as real-time correction values for a $5^\circ \times 5^\circ$ grid covering the Indian subcontinent. Both data sets are determined by the IRNSS Navigation Center from the monitoring stations distributed across the service area (Sect. 11.3.4).

11.3.3 Spacecraft

Platform

The IRNSS spacecraft is configured around the INSAT-1000 (I1K) platform, an indigenous 1 t class bus for three-axis stabilized spacecraft. The I1K bus has been selected because the weight of the payload and all the subsystems required for IRNSS can be accommodated in the structure. It offers a favorable ratio of structural weight to payload weight and volume, and meets the propulsion requirement of the satellites. The I1K bus also falls within the payload capacity of the PSLV (Polar Satellite Launch Vehicle), the workhorse of Indian space program. All the satellites for GEO and IGSO will be similar. The common design has been chosen in order to manufacture the satellite in a production mode to enable the deployment of constellation in a fast-track mode. Key parameters of the IRNSS spacecraft are summarized in Table 11.11. The stowed view of the IRNSS spacecraft is shown in Figs. 11.20 and 11.21.

The power system comprises the solar panels used to generate the required power, the batteries for storage and, eclipse support, and the electronics to manage, control, and distribute the power throughout the spacecraft. The solar cells generate a total power of around 1600 W to support a payload power requirement of 900 W.

The attitude and orbit control subsystem (AOCS) of IRNSS is configured as a three-axis stabilized zero momentum system with reaction wheels to provide a stable platform for the navigation application. The AOCS together with the propulsion subsystem facilitates the transfer orbit maneuvers, station-keeping maneuvers, and the fine adjustment of the spacecraft attitude. The star sensors provide the orientation data and the dynamically tuned gyroscope (DTG) delivers the body rates data to the AOCS, which computes the necessary control torque commands for the actuators. Other than common MEO GNSS satellites, the IRNSS spacecraft do not employ a nadir orientation, but continuously steer the yaw axis (and thus the antenna boresight) toward a point at 83° E longitude and 5° N latitude on the

surface of the Earth. In this way, an optimum signal coverage is obtained for the envisaged service area. The solar panels are oriented to generate maximum power by maintaining the Sun direction normal to the panel surface.

The thermal control system provides a benign thermal environment to the various subsystem elements of the spacecraft. A judicious combination of radiators and insulators is used to achieve the thermal balance. While conventional thermal control elements are used for the majority of the subsystems, special thermal control schemes have been designed and implemented for temperature-sensitive elements such as atomic clocks, clock-monitoring units, and the corner cube retro reflector.

The IRNSS satellites are equipped with a 440 N boost motor that is used to reach the desired mission orbit after separation from the launcher and deployment into the elliptical transfer orbit [11.80]. The required fuel contributes the dominant fraction of the wet spacecraft mass at lift-off. Further to the boost motor, a set of 12 22 N thruster is accommodated for fine orbit acquisition and maintenance maneuvers.

Payload

The IRNSS satellites have two primary types of payloads, viz. the navigation payload and the ranging payload. The navigation payload of IRNSS covers the generation of navigation signals in the S- and L5-bands and their transmission to the users. The ranging payload, in contrast, supports two-way ranging of the IRNSS spacecraft from dedicated ground stations as part of the mission operations. As a third payload, the IRNSS satellites carry a retroreflector array for SLR.

Navigation Payload. The architecture of the IRNSS navigation payload is illustrated in Fig. 11.22. It comprises the following major subsystems:

- The RAFSs
- The atomic clock monitoring unit (ACMU)
- The clock distribution unit (CDU)
- The navigation signal generation unit (NSGU)
- The modulation and up-converter unit
- The high-power traveling-wave tube amplifier (TWTa)
- Output filters, and
- The dual-band array antenna.

A redundant set of three atomic clocks provides the basis for the onboard time and frequency generation. The RAFSs selected for the IRNSS satellites have been manufactured by Spectracom, Switzerland, which has earlier provided similar RAFS for the Galileo and



Fig. 11.20 IRNSS spacecraft in clean room (courtesy of ISRO)

Table 11.11 IRNSS spacecraft characteristics [11.90]

Parameter	Value
Dry mass	614 kg
Lift off mass	1425 kg
Physical dimension	1.58 m × 1.5 m × 1.5 m
Power generation	Two solar panels generating 1660 W, one lithium ion battery with 90 Ah capacity
Propulsion	440 N liquid apogee motor and 12 22 N thrusters
Control system	Zero momentum system; Sensors: Sun sensors, star sensors, gyroscopes; Actuators: reaction wheels, magnetic torquers, 22 N thrusters
Mission life	10 years

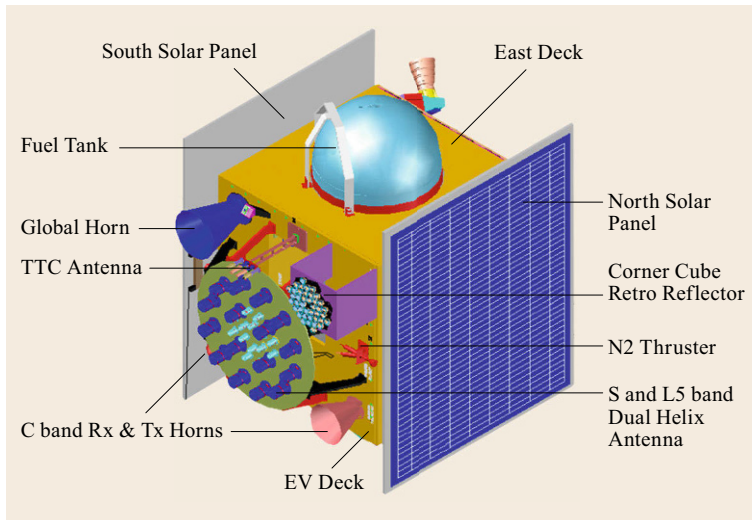


Fig. 11.21 IRNSS satellite – stowed view (courtesy of ISRO)

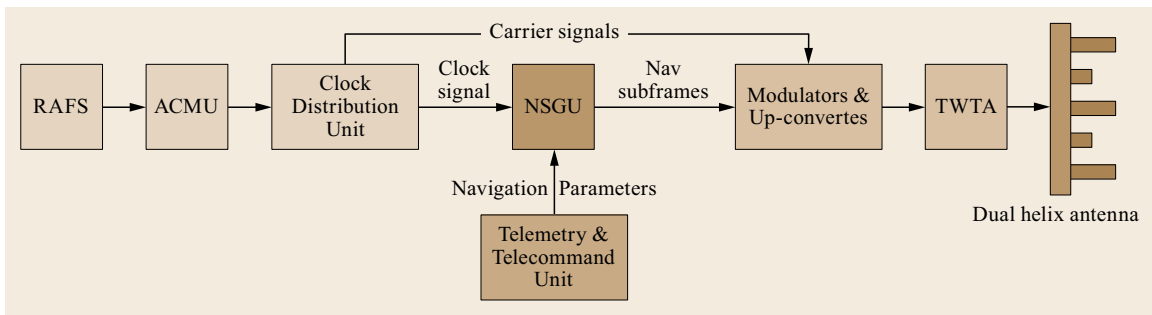


Fig. 11.22 Navigation payload block diagram

BeiDou programs. They offer very high stability with an ADEV of better than $0.5 \cdot 10^{-12}(\tau/s)^{-1/2}$ over time scales from $\tau = 1$ to 10000 s

Within the ACMU, the frequencies of individual frequency standards are compared and monitored. In addition, the fundamental frequency $f_0 = 10.23$ MHz is generated from the native 10 MHz output of the active RAFS. All carrier frequencies and chipping rates required for the navigation signal are rational multiples of 10.23 MHz (e.g., $f_{L5} = 115f_0$ and $R_{c_sps} = 0.1f_0$) and are coherently generated from the ACMU output through frequency multiplication or division.

The NSGU receives the navigation parameters computed on ground through the telemetry and telecommand unit and stores them in the NSGU memory. The stored data are time stamped, encoded, formatted, and combined into individual messages and subframes. The resulting broadcast navigation data are modulo-2 added to the satellite-specific PRN code and modulated on the carrier. Finally, the modulator output is up-converted, amplified to the required power level, and transmitted through the antenna.

To obtain a gain pattern with the desired shape and directivity, the IRNSS antenna is made up of an array of phase-coherently connected helix antenna elements. However, distinct elements are used for the L5-band (1176.45 MHz) and the S-band (2492.028 MHz), which differ by more than a factor of 2 in frequency and require different physical dimensions of the respective helix elements. As illustrated in Fig. 11.21, the antenna comprises 16 and 18 short axial helix elements for the L5- and S-band, respectively. The combined antenna has been designed such that the phase centers of both bands lie on the same axis, which also meets the overall antenna RF performance requirements. For thermal protection, the antenna array is covered with an insulation layer.

Ranging Payload. The IRNSS satellite has an independent C-band bent-pipe transponder for ranging. A code division multiple access (CDMA) modulation is employed for the ranging signal. This enables concurrent two-way ranging of the IRNSS satellites from up to four ground-based ranging stations [11.91].

The ranging payload comprises distinct C-band horn antennas for the uplink and downlink (Fig. 11.21), a pre-select filter, the receiver subsystem, the solid state power amplifier (SSPA), and the output band-pass filter. The narrow band transponder has a bandwidth of 25 MHz.

The CDMA ranging stations (IRCDR) located at different places in India perform the two-way ranging of the satellite and facilitate precise orbit determination of the IRNSS satellites. The two-way ranging also helps to validate the one-way range measurements performed by the reference stations (IRNSS Range and Integrity Monitoring Stations (IRIMS)).

Corner Cube Retro Reflectors. Corner cube retroreflectors (CCRRs) are placed on the IRNSS spacecraft to enable precise laser ranging. The retroreflector array has been developed in India is composed of 40 uncoated Suprasil-311 quartz prisms with a circular aperture of 38 mm diameter (Fig. 11.23).

The IRNSS satellites are routinely tracked by European, Australian, and Asian stations under the coordination of the ILRS [11.62]. The resulting observations offer a cm-level precision and are mainly used for the validation of the IRNSS precise orbit determination obtained from radiometric one- and two-way tracking. In addition, they can also be used for standalone orbit determination, even though the temporal coverage is substantially less dense than that of radiometric tracking data [11.92–94].

Launch and Orbit Injection

The IRNSS satellites are launched from the Satish Dhawan Space Center (SDSC) in Shriharikota at the East Coast of India using the XL version of the well-proven PSLV (Fig. 11.24). With a total of four stages, the PSLV has an overall mass of 320 t and a height of

44 m. Solid propellant is employed for the first stage (which includes a core stage and six strap-on boosters) as well the third stage, whereas liquid fuel is used in stages 2 and 4 [11.95].

Following the burnout of the fourth stage near 20 min after lift-off, the IRNSS satellite is deployed into a transfer orbit of about 18° inclination, which reaches from a perigee height near 300 km to a peak altitude of 20 700 km at apogee. To achieve the desired geosynchronous orbit, the apogee is first raised to roughly 36 000 km using two burns of the satellites 440 N boost motor. Subsequently, the perigee is raised to the same altitude using a series of three consecutive apogee boot maneuvers. These are also used to adjust the orbital inclination to the desired target value. Details of the mission design and maneuver planning are provided in [11.80] for the example of the IRNSS-1A satellite.

11.3.4 Ground Segment

The IRNSS/NavIC ground segment comprises various infrastructure components supporting the spacecraft and mission operations [11.90]. It includes:

- The satellite control facility
- The navigation centre, and the
- The network timing facility, as well as
- Range and integrity monitoring stations
- CDMA ranging stations, and
- The data communication network.

These are complemented by SLR stations of external institutions providing supplementary tracking of the IRNSS satellites on a best effort basis. The location of the Indian ground segment sites is illustrated in Fig. 11.25.

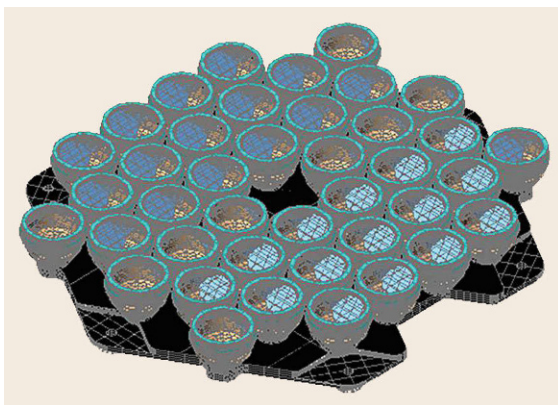


Fig. 11.23 Laser retroreflector array of the IRNSS satellites (courtesy of ISRO/ISTRAC)



Fig. 11.24 The PSLV used to carry the first IRNSS satellite into orbit (courtesy of ISRO)

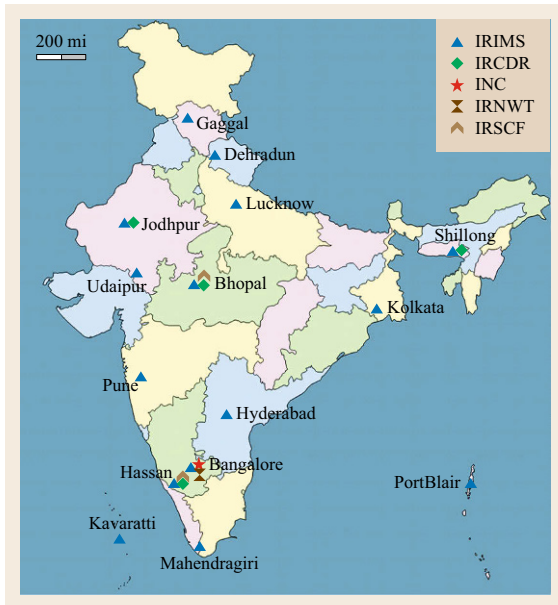


Fig. 11.25 Map of the IRNSS ground segment

IRNSS Satellite Control Facility

TT&C operations of the IRNSS satellites are carried out from the IRNSS Satellite Control Facility (IRSCF; Fig. 11.26). It comprises two Satellite Control Centres (SCC) and Spacecraft Control Earth Stations (SCES). For redundancy purposes, the main SSC at Hassan is complemented by a second SSC at Bhopal.

The Spacecraft Control Centre (SCC) consists of all the computers, servers, encoders, the monitoring and command software, etc., used to conduct the operations of the spacecraft platform. It monitors the telemetry signals, generates the telecommands, encodes them and forward them to SCES. These consist of several full coverage and full motion antennas, which are used to track the IRNSS spacecraft, to acquire their telemetry, and to uplink telecommands.

In addition to its regular (TT&C) operation, the IRSCF also receives the navigation parameters generated by the navigation software from the INC and uplinks them to the spacecraft.

As a GEO-IGSO combination of satellites has been chosen for the IRNSS space segment, regular station-keeping operations are required roughly once per month [11.94] to maintain the specified orbits and ground tracks. The station-keeping maneuvers cause a short outage of service for a given satellite. On completion of the station-keeping maneuver, the ranging and orbit determination of the satellite are again carried out and the new set of orbit and clock parameters are uplinked before the satellite is again declared op-



Fig. 11.26 IRNSS spacecraft control facility Hassan (courtesy of ISRO)

erational. The processes are highly automated and the outage time is kept minimal.

IRNSS Navigation Centre

The INC at Byalalu near Bangalore is responsible for ensuring the navigation operations of IRNSS. It acquires and assimilates pseudorange and carrier-phase observations from the monitoring stations as well as two-way CDMA and satellite laser ranging measurements from the respective stations.

The navigation processing unit within the INC performs various functions such as range data processing, orbit determination, validation, and prediction satellite, station and system time computation, clock parameter estimation and prediction, as well as the ionospheric modeling. Based on these computations, the navigation software generates the primary and secondary navigation parameters, which are subsequently uplinked to the IRNSS satellites. The INC also performs remote monitoring and control of all distributed facilities from a centralized location.

The orbit determination and prediction of IRNSS use state-of-the art concepts for the trajectory and measurement modeling. Using different types of observations, a 3-D RMS consistency at the 10–20 m level has been demonstrated for individual solutions. A dedicated radiation pressure model with a priori box-cylinder-wing model and three adjustable parameters [11.96] is employed in the orbit computation to properly account for the specific spacecraft properties. The orbit determination system makes use of a batch least-squares estimator for routine operations but switches to an extended Kalman filter (EKF) for post-maneuver arcs or in the case of unexpected clock events to achieve the best performance and to minimize the associated outages [11.92, 93]. When utilizing the EKF mode, the ephemeris updated interval is reduced from a nominal value of 2 h down to 15 min. Users may distinguish the EKF mode from the *Issue of Data Ephemeris and Clock* (IODEC) parameter in the ephemeris message, which is

nominally in the range of 0–11, but assumes values of 160–254 after a maneuver [11.94].

Generally, the orbit and clock parameters are predicted in advance for 24 h and uplinked to the spacecraft. In order to cater for contingency situations resulting in a communication breakage from ground system to the spacecraft, AutoNav data are also uplinked to the spacecraft. The IRNSS satellites store 7 days of AutoNav data sets with ephemeris and clock parameters and support a subsequent broadcast of primary navigation parameters.

In order to mitigate the ionospheric errors for a single frequency user, IRNSS utilizes a specially devised grid-based ionospheric model in addition to the conventional Klobuchar-like model. The ionospheric grid parameters computed by the navigation software and uplinked to the spacecraft form the major part of the secondary navigation data. The time offsets w.r.t. other GNSS service providers are also computed and uplinked as a part of secondary navigation parameters. These time offset parameters facilitate time correlation between different systems in multi-GNSS receivers.

IRNSS Network Timing Facility

The IRNSS Network Timing (IRNWT) Facility at Byalalu has been established with the objective to generate and disseminate IRNSS system time as an independent IRNSS time scale with the required accuracy, stability and traceability to UTC. It covers the following functions [11.97]:

- Precise timekeeping to support the navigation mission
- Steering of IRNSS System Time (IRNSST) toward international atomic time (TAI) and provision of IRNSST-UTC time information to the users.

The IRNSS system time is realized by an ensemble of atomic clocks such as active hydrogen-maser (AHM) and cesium standards, with appropriate measurement equipments and timing algorithms [11.98]. The resulting time scale is steered to International Atomic Time (TAI). TWSTFT as well as GNSS common-view time transfer methods are adopted for time steering. IRNSST will be steered to TAI by linking with UTC(k) labs such that the difference between TAI and IRNSSTT is maintained within 50 ns (2σ) at any yearly time interval.

IRNSS Range and Integrity Monitoring Stations

The IRIMS are located at well-surveyed locations covering the IRNSS service area. They perform the one-way range measurements, which form the primary set of observations for orbit determination and clock offset estimation of the IRNSS spacecraft. The IRIMS house

the IRNSS reference receivers that continuously track the navigation signals from the IRNSS constellation and transmit the pseudorange and carrier-phase measurements to the NC for further processing. IRIMS also aid the integrity determination of the IRNSS constellation and are used for determining the ionospheric and tropospheric delays as well as other biases in the IRNSS signals.

The NovAtel GIII receivers (Fig. 11.27) used in the IRNSS monitoring stations are based on the design of the reference station receivers for the United States' Wide Area Augmentation System, but have been adapted to tracking of the IRNSS SPS signals in the L5- and S-band.

IRNSS CDMA Ranging Stations

Two-way ranging [11.91] is used for orbit determination of IRNSS satellites, in addition to one-way range measurements from IRIMS. It is carried out through the IRNSS CDMA Ranging (IRCDR) stations placed at widely separated locations to meet the required orbit determination accuracies. Each station has a full motion antenna system, utilized in time-shared mode, to perform ranging measurements for all satellites in the constellation. The two-way ranging data are processed at INC to estimate the orbit by which the one-way range measurements are validated.

Satellite Laser Ranging Stations

SLR provides precise distance measurements for the calibration of the IRIMS one-way ranging and CDMA two-way ranging of IRNSS, and for validating the orbit estimations obtained from these measurement systems. Laser ranging of the IRNSS satellites is accomplished by measuring the round-trip time of a laser pulse reflected from a retroreflector array mounted on the Earth



Fig. 11.27 GIII reference receiver for the IRIMS (courtesy of NovAtel)

facing side of the satellite. IRNSS takes the support of the ILRS [11.62] to perform the laser ranging of its satellites in a campaign mode using stations located across the globe.

IRNSS Data Communication Network

The IRNSS Data Communication Network (IRDCN) is a high-availability communications network consisting of redundant terrestrial and satellite links. The IRDCN enables the digital communication between the IRIMS, ICRDR, IRNWT, IRSCF, and INC with the highest order of reliability and availability.

11.3.5 System Performance

Generally speaking, the user positioning accuracy can be described by the product of the user equivalent range error (UERE) and the Position Dilution of Precision (PDOP, [11.99]). Following [11.100], position dilution of precision (PDOP) values of about 3 can be achieved with the full constellation of seven IRNSS satellites over the Indian subcontinent.

The UERE is defined as the RMS of the pseudorange measurements and modeling errors experienced by a user receiver in the navigation process. These comprise the signal-in-space range error (SISRE) describing the impact of orbit and clock errors as well as user-specific errors such as receiver noise and multipath or unmodeled atmospheric path delays.

To assess the impact of the navigation message and to quantify the achieved SISRE, the difference between the smoothed pseudorange observations of the IRNSS monitoring stations and their modeled values for the given location is continuously monitored by the INC. As illustrated in Fig. 11.28 for the first three satellites of the IRNSS constellation, meter-level line-of-sight errors are obtained during normal operations, which yield an SISRE of 1–2 m in accord with the user range accuracy (URA) values reported in the navigation message during these phases. Increased line-of-sight errors of up to 5 m may, however, be encountered after orbit keeping maneuvers [11.92].

A very first standalone navigation fix was achieved in April 2015 shortly after the activation of the IRNSS-

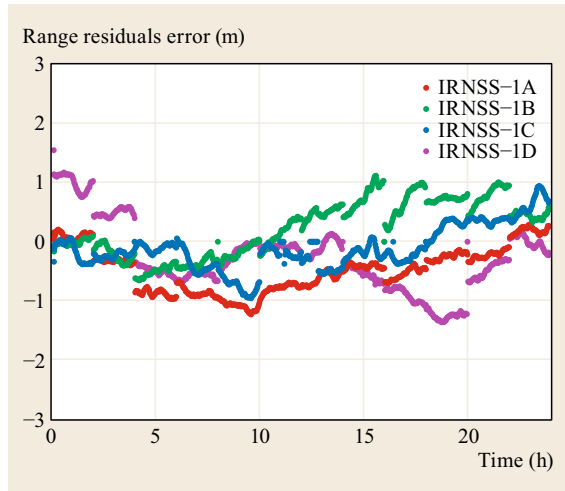


Fig. 11.28 Difference of observed and modeled pseudoranges for IRNSS-1A, -1B, -1C, and -1D w.r.t. a monitoring station over a maneuver-free 24 h period (4 Sep. 2015)

1D satellite. As discussed in [11.101], 3-D positioning accuracies of 10 m or better could be obtained with the four satellite constellation except during selected periods of highly unfavorable PDOP. With the launch of the seventh satellite IRNSS-1G, positioning accuracy better than 10 m is obtained at any time over the Indian Region. The initial experiments also confirmed the quality of the ionospheric corrections and demonstrated that single-frequency users with corrections will be able to achieve a similar performance as dual-frequency L5/S-band users.

The interoperability of the IRNSS L5 signal with other global and regional navigation satellite systems has been successfully demonstrated in [11.102]. Aside from evidencing a consistent measurement quality with that of GPS, Galileo and QZSS L5/E5a signals, the IRNSS observations were, for the first time, used in a combined multi-GNSS relative position solution together with the three other constellations. Compared to L1 signals, the longer L5 wavelength facilitates ambiguity resolution and the incorporation of IRNSS offers improved geometry and additional robustness for the obtained solutions.

References

- 11.1 C. Carnebianca: Regional to global satellite based navigation systems, IEEE PLANS'88, Orlando (1988) pp. 25–33
- 11.2 J.R. Wertz, W.J. Larson: *Space Mission Analysis and Design*, 3rd edn. (Microcosm, Torrance 1999) pp. 143–144
- 11.3 R.D. Briskman: Radio Determination Satellite Service, Proc. IEEE **78**(7), 1096–1106 (1990)
- 11.4 R.D. Briskman, R.J. Prevaux: S-DARS broadcast from inclined, elliptical orbits, Acta Astronaut. **54**(7), 503–518 (2004)
- 11.5 M. Tanaka, K. Kimura, E. Morikawa, A. Miura, S. Kawase, S. Yamamoto, H. Wakana: Application technique of figure-8 satellites system, Technical Report SAT 99(45), 55–62 (Institute of Electronics, Information and Communication Engineers) in Japanese
- 11.6 H.D. Takahashi: Japanese regional navigation satellite system "The JRANS Concept", J. Glob. Position. Syst. **3**(1/2), 259–264 (2004)
- 11.7 S. Kogure, M. Kishimoto, M. Sawabe: Future expansion from QZSS to regional satellite navigation system, ION NTM, San Diego (ION, Virginia 2007) pp. 455–460
- 11.8 J. Spilker: Satellite constellation and geometric dilution of precision. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington 1996) pp. 177–208
- 11.9 L. Ma, S. Li: Mathematical aspects for RNSS constellation with IGSO satellites, Earth Sci. Res. **3**(2), 66–71 (2014)
- 11.10 I. Kawano, M. Mokuno, S. Kogure, M. Kishimoto: Japanese experimental GPS augmentation using Quasi-Zenith Satellite System (QZSS), ION GNSS, Long Beach (ION, Virginia 2004) pp. 175–181
- 11.11 Y. Murai: Project overview of the Quasi-Zenith Satellite System, Proc. ION GNSS+, Tampa (ION, Virginia 2015) pp. 1291–1332
- 11.12 A. Matsumoto: Status update on the Quasi-Zenith Satellite System (QZSS), 9th Meet. Int. Comm. GNSS (ICG), Prague (UNOOSA, Vienna 2014) pp. 1–18
- 11.13 Service overview on the Quazi-Zenith Satellite System (QZSS) web site, <http://qzss.go.jp/en/overview/services/>
- 11.14 Japan Aerospace Exploration Agency: Quasi-Zenith Satellite System navigation service interface specification for QZSS, IS-QZSS, VI.6 (JAXA, 2014)
- 11.15 S. Kogure, I. Kawano: GPS augmentation and complement using Quasi-Zenith Satellite System (QZSS), AIAA 2003–2416, Proc. 21st AIAA Int. Commun. Satell. Syst. Conf. Exhib., Yokohama (AIAA, Reston 2003) pp. 1–10
- 11.16 K. Kimura, M. Tanaka: Required velocity increment for formation keeping of inclined geosynchronous constellations, Proc. 51st Int. Astronaut. Cong., Rio de Janeiro (IAF, Paris 2000)
- 11.17 Y. Murai: Project overview Quasi-Zenith Satellite System, Symp. Commer. Appl. Global Navig. Satell. Syst., Vienna (UNOOSA, Vienna 2014) pp. 1–33
- 11.18 M. Saito, J. Takiguchi, T. Okamoto: Establishment of regional navigation satellite system utilizing quasi-zenith satellite system, Mitsubishi Electr. Adv. Mag. **147**, 1–6 (2014)
- 11.19 Quasi-Zenith Satellite System Interface Specification – Satellite Positioning, Navigation and Timing Service, IS-QZSS-PNT-001, Draft 12 July 2016 (Cabinet Office, 2016)
- 11.20 Quasi-Zenith Satellite System Interface Specification – Centimeter Level Augmentation Service, IS-QZSSL6-001, Draft 12 July 2016 (Cabinet Office, 2016)
- 11.21 Quasi-Zenith Satellite System Interface Specification – Positioning Technology Verification Service, IS-QZSS-TV-001, Draft 12 July 2016 (Cabinet Office, 2016)
- 11.22 Navstar GPS Space Segment / Navigation User Segment Interfaces, Interface Specification, IS-GPS-200H, 24 Sep. 2013 (Global Positioning Systems Directorate, 2013)
- 11.23 Navstar GPS Space Segment / User Segment L5 Interfaces, Interface Specification, IS-GPS-705D, 24 Sep. 2013 (Global Positioning Systems Directorate, 2013)
- 11.24 Navstar GPS Space Segment / User Segment L1C Interfaces, Interface Specification, IS-GPS-800D, 24 Sep. 2013 (Global Positioning Systems Directorate, 2013)
- 11.25 L1 C/A PRN Code Assignments; US Air Force, Los Angeles Air Force Base, 6 Jan. 2016. <http://www.losangeles.af.mil/About-Us/Fact-Sheets/Article/734549/gps-prn-assignment>
- 11.26 J.W. Betz: Binary offset carrier modulations for radionavigation, Navigation **48**(4), 227–246 (2001)
- 11.27 J.W. Betz, M.A. Blanco, Ch.R. Cahn, Ph.A. Dafesh, Ch.J. Hegarty, K.W. Hudnut, V. Kasemsri, R. Keegan, K. Kovach, L.S. Lenahan, H.H. Ma, J.J. Rushanan, D. Sklar, T.A. Stansell, C.C. Wang, S.K. Yi: Description of the L1C signal, ION GNSS, Fort Worth (ION, Virginia 2006) pp. 2080–2209
- 11.28 H. Maeda: System Research on The Quasi-Zenith Satellites System (in Japanese), Ph.D. Thesis (Tokyo University of Marine Science and Technology, Tokyo 2007)
- 11.29 Technical Working Group Report to the U.S.-Japan GPS Plenary, (GPS-QZSS Technical Working Group, 18 Jan. 2012) <http://www.gps.gov/policy/cooperation/japan/2012-joint-announcement/TWG-report.pdf>
- 11.30 T. Sakai, H. Yamada, S. Fukushima, K. Ito: Generation and evaluation of QZSS L1-SAIF ephemeris information, ION GNSS, Portland (ION, Virginia 2011) pp. 1277–1287
- 11.31 S. Thoeleert, S. Erker, J. Furthner, M. Meurer: Lat-est signal in space analysis of GPS IIF, COMPASS and QZSS, NAVITEC'2010, Noordwijk (ESA, Noordwijk 2010) pp. 1–8
- 11.32 RTCA D0229D Change 1: Minimum Operational Performance Standards for Global Positioning Sys-

- tem/Wide Area Augmentation System Airborne Equipment (RTCA, Feb. 2013)
- 11.33 T. Sakai, S. Fukushima, N. Takeichi, K. Ito: Implementation of the QZSS L1-SAIF message generator, ION NTM, San Diego (ION, Virginia 2008) pp. 464–476
- 11.34 T. Sakai, S. Fukushima, K. Ito: QZSS L1-SAIF Initial Experiment Results, ION ITM, San Diego (ION, Virginia 2011) pp. 1133–1142
- 11.35 R. Iwama, H. Soga, K. Odagawa, Y. Masuda, T. Osawa, A. Ito, M. Matsumoto: Operation of sub-meter class augmentation system and demonstration experiments with Quasi-Zenith Satellite “MICHIBIKI”, ION ITM, Newport Beach (ION, Virginia 2012) pp. 1295–1301
- 11.36 T. Sakai, H. Yamada, K. Ito: Ranging quality of QZSS L1-SAIF signal, ION ITM, Newport Beach (ION, Virginia 2012) pp. 1255–1264
- 11.37 S. Choy, K. Harima, Y. Li, M. Choudhury, C. Rizos, Y. Wakabayashi, S. Kogure: GPS precise point positioning with the Japanese Quasi-Zenith Satellite System LEX augmentation corrections, *J. Navig.* **68**(4), 769–783 (2015)
- 11.38 T. Kasami: Weight distribution formula for some class of cyclic codes, Technical Report R285, 1–24 (University of Illinois, Urbana-Champaign 1966)
- 11.39 S. Kogure: Evaluation of QZS-1 LEX signal, 7th Meet. Int. Comm. GNSS (ICG), Work. Group B, Beijing (UN-OOSA, Vienna 2012) pp. 1–9
- 11.40 S. Choy, K. Harima, Y. Li, Y. Wakabayashi, H. Tateshita, S. Kogure, C. Rizos: Real-time precise point positioning utilising the Japanese quasi-zenith satellite system (QZSS) LEX corrections, *Proc. IGSSS Symp., Surfers Paradise* (IGSSS Society, Tweed Heads 2013) pp. 1–15
- 11.41 A. Garcia-Pena, D. Salos, O. Julien, L. Ries, T. Grellier: Analysis of the use of CSK for future GNSS Signals, ION GNSS, Nashville (ION, Virginia 2013) pp. 1461–1479
- 11.42 Y. Hatanaka, Y. Kuroishi, H. Munekane, A. Wada: Development of a GPS Augmentation Technique, *Proc. Int. Symp. GPS/GNSS – Toward New Era Position. Technol.*, Tokyo (GPS/GNSS Society Japan, 2008) pp. 1097–1103
- 11.43 M. Saito, K. Asari: Centimeter-class Augmentation System (CMAS), *Proc. ION GNSS*, Nashville (ION, Virginia 2012) pp. 3354–3365
- 11.44 RTCM Standard 10403.2: Differential GNSS Services, Version 3 with Amendment 2, 7 Nov. 2013 (RTCM, Arlington, VA 2013)
- 11.45 M. Schmitz: RTCM state space representation messages, status and plans, PPP-RTK Open Stand. Symp., Frankfurt (2012) pp. 1–31
- 11.46 M. Caissy, L. Agrotis, G. Weber, M. Hernandez-Pajares, U. Hugentobler: Coming soon – The international GNSS real-time service, *GPS World* **23**(6), 52 (2012)
- 11.47 M. Saito, Y. Sato, M. Miya, M. Shima, Y. Omura, J. Takiguchi, K. Asari: Centimeter-class Augmentation System Utilizing Quasi-Zenith Satellite, ION GNSS, Portland (ION, Virginia 2011) pp. 1243–1253
- 11.48 T. Suzuki, N. Kubo, T. Takasu: Evaluation of precise point positioning using MADOCA-LEX via Quasi-Zenith Satellite System, ION ITM, San Diego (ION, Virginia 2014) pp. 460–470
- 11.49 M. Homma, S. Yoshimoto, N. Natori, Y. Tsutsumi: Engineering Test Satellite-8 for mobile communications and navigation experiment, *Proc. 51st Int. Astronaut. Cong.*, Rio de Janeiro (IAF, Paris 2000)
- 11.50 N. Inaba, A. Matsumoto, H. Hase, S. Kogure, M. Sawabe, K. Terada: Design concept of Quasi Zenith Satellite System, *Acta Astronaut.* **65**(7), 1068–1075 (2009)
- 11.51 Y. Ishijima, N. Inaba, A. Matsumoto, K. Terada, H. Yonechi, H. Ebisutani, S. Ukava, T. Okamoto: Design and development of the first quasi-zenith satellite attitude and orbit control system, *IEEE Aerosp. Conf.*, Big Sky (2009) pp. 1–8, doi:10.1109/AERO.2009.4839537
- 11.52 O. Montenbruck, R. Schmid, F. Mercier, P. Steigenberger, C. Noll, R. Fatkulin, S. Kogure, S. Ganeshan: GNSS satellite geometry and attitude models, *Adv. Sp. Res.* **56**(6), 1015–1029 (2015)
- 11.53 A. Hauschild, P. Steigenberger, C. Rodriguez-Solano: QZS-1 Yaw attitude estimation based on measurements from the CONGO network, *Navigation* **59**(3), 237–248 (2012)
- 11.54 H. Noda, S. Kogure, M. Kishimoto, H. Soga, T. Moriguchi, T. Furubayashi: Development of the quasi-zenith satellite system and high-accuracy positioning experiment system flight model, *NEC Tech. J.* **5**(4), 93–97 (2010)
- 11.55 T. Obara, S. Furuhashi, H. Matsumoto: Overview of initial observation data of technical data acquisition equipments on the first Quasi-Zenith Satellite, 2011-r-58, *Proc. 28th Int. Symp. Space Technol. Sci. (ISTS)*, Okinawa (ISTS, Tokyo 2011) pp. 1–4
- 11.56 S. Hama, Y. Takahashi, K. Kimura, H. Ito, J. Amagai: Quasi-Zenith Satellite System (QZSS) Project, *J. Natl. Inst. Inf. Commun. Technol.* **57**(3/4), 289–296 (2010)
- 11.57 M. Nakamura, Y. Takahashi, J. Amagai, T. Gotoh, M. Fujieda, R. Tabuchi, S. Hama, Y. Yahagi, T. Takahashi, S. Horiuchi: Time comparison experiments between the QZS-1 and its time management station, *Navigation* **60**(4), 319–324 (2013)
- 11.58 O. Montenbruck, P. Steigenberger, E. Schönmann, A. Hauschild, U. Hugentobler, R. Dach, M. Becker: Flight characterization of new generation GNSS satellite clocks, *Navigation* **59**(4), 291–302 (2012)
- 11.59 H. Ito, T. Morikawa, S. Hama: Development and performance evaluation of spaceborne hydrogen maser atomic clock in NICT, ION NTM, San Diego (ION, Virginia 2007) pp. 452–454
- 11.60 T. Iwata, T. Matsuzawa, K. Machita, T. Kawauchi, S. Ota, Y. Fukuhara, T. Hiroshima, K. Tokita, T. Takahashi, S. Horiuchi, Y. Takahashi: Demonstration experiments of a remote synchronization system of an onboard crystal oscillator using “MICHIBIKI”, *Navigation* **60**(2), 133–142 (2013)

- 11.61 S. Nakamura: Impact of SLR tracking on QZSS, Proc. Int. Tech. Workshop SLR Track. GNSS Constellations, Metsovo, ed. by E. Pavlis (ILRS, Greenbelt 2009) pp. 68–92
- 11.62 M.R. Pearlman, J.J. Degnan, J.M. Bosworth: The International Laser Ranging Service, Adv. Space Res. **30**(2), 135–143 (2002)
- 11.63 O. Montenbruck, P. Steigenberger, G. Kirchner: GNSS satellite orbit validation using satellite laser ranging, Proc. 18th Int. Workshop Laser Ranging, Fujiyoshida (ILRS, Greenbelt 2013) pp. 13–0209
- 11.64 K. Akiyama, T. Otsubo: Accuracy evaluation of QZS-1 orbit solutions with Satellite Laser Ranging, Proc. ILRS Tech. Laser Workshop Satell., Lunar Planet. Laser Ranging: Charact. Space Segment, Frascati (ILRS, Greenbelt 2012)
- 11.65 N. Inaba, H. Hase, H. Miyamoto, Y. Ishijima, S. Kawakita: A satellite simulator and model based operations in Quasi-Zenith Satellite System, AIAA Model. Simul. Conf., AIAA-2009-5813, Chicago (AIAA, Reston 2009) pp. 1–16
- 11.66 H. Miyamoto, M. Kishimoto, E. Myojin, S. Kogure: Model-based design of Ground Segment for Quasi-Zenith Satellite System, Proc. SpaceOps 2012 Conf., Stockholm (AIAA, Reston 2012) pp. 1–7
- 11.67 M. Nakamura, S. Hama, Y. Takahashi, J. Amagai, T. Gotoh, M. Fujieda, R. Tabuchi, M. Aida, I. Nakazawa, T. Hobiger, T. Takahashi, S. Horiuchi: Time management system of the QZSS and time comparison experiments, AIAA 2011-8067, 29th AIAA Int. Commun. Satell. Syst. Conf. (ICSSC-2011), Nara (AIAA, Reston 2011) pp. 534–538
- 11.68 N. Kajiwaru, Y. Yamamoto, M. Sawabe, S. Kogure, T. Tsuruta, M. Kishimoto, Y. Kawaguchi, T. Shibata: Overview of precise orbit and clock estimation for Quasi-Zenith Satellite System and simulation results, 2009-d-35, Proc. 27th Int. Symp. Space Technol. Sci. (ISTS), Tsukuba (ISTS, Tokyo 2009) pp. 1–6
- 11.69 S. Matsumura, M. Murakami, T. Imakiire: Concept of the new Japanese geodetic system, Bull. Geogr. Surv. Inst. **51**, 1–9 (2004)
- 11.70 J.A. Klobuchar: Ionospheric time-delay algorithm for single-frequency GPS users, IEEE Trans. Aerosp. Electron. Syst. AES-2 **3**(3), 325–331 (1987)
- 11.71 E.M. Soop: *Handbook of Geostationary Orbits* (Kluwer Academic, Dordrecht 1994)
- 11.72 Notice Advisory to QZSS Users (JAXA), <http://qz-vision.jaxa.jp/USE/en/naqu>
- 11.73 T. Sawamura, T. Takahashi, T. Moriguchi, K. Ohara, H. Noda, S. Kogure, M. Kishimoto: Performance of QZSS (Quasi-Zenith Satellite System) and L-Band Navigation Payload, ION GNSS, Nashville (ION, Virginia 2012) pp. 1228–1254
- 11.74 E. Kishimoto, M. Myojin, S. Kogure, H. Noda, K. Terada: QZSS On Orbit Technical Verification Results, ION GNSS, Portland (ION, Virginia 2011) pp. 1206–1211
- 11.75 JAXA: “QZ-vision” Experiment Results SIS-URE, http://qz-vision.jaxa.jp/USE/en/exp_results_report
- 11.76 O. Montenbruck, P. Steigenberger, A. Hauschild: Broadcast versus precise ephemerides: A Multi-GNSS perspective, GPS Solut. **19**(2), 321–333 (2015)
- 11.77 F. Gonzalez, P. Waller: GNSS clock performance analysis using one-way carrier phase and network methods, 39th Annu. Precise Time Time Interval (PTTI) Meet., Long Beach (ION, Virginia 2007) pp. 403–414
- 11.78 P. Steigenberger, A. Hauschild, O. Montenbruck, C. Rodriguez-Solano, U. Hugentobler: Orbit and clock determination of QZS-1 based on the CONGO network, Navigation **60**(1), 31–40 (2013)
- 11.79 A.S. Ganeshan, S.C. Rathnakara, R. Gupta, A.K. Jain: Indian Regional Navigation Satellite System (IRNSS) Concept, J. Spacecr. Technol. **15**(2), 19–23 (2005)
- 11.80 B.S. Kiran, S. Singh: Mission design and analysis for IRNSS-1A, Proc. 65th Int. Astronaut. Congr., Toronto (IAF, Paris 2000) pp. 1–12
- 11.81 P. Majithiya, K. Khatrri, J.K. Hota: Indian Regional Navigation Satellite System – Correction parameters for timing group delays, Inside GNSS **6**(1), 40–46 (2011)
- 11.82 S. Thoeleert, O. Montenbruck, M. Meurer: IRNSS-1A – Signal and clock characterization of the Indian Regional Navigation System, GPS Solutions **18**(1), 147–152 (2014)
- 11.83 S.B. Sekar, S. Sengupta, K. Bandyopadhyay: Spectral compatibility of BOC(5,2) modulation with existing GNSS signals, Proc. IEEE/ION PLANS 2012, Myrtle Beach (2012) pp. 886–890
- 11.84 Indian Regional Navigation Satellite System – Signal In Space ICD for Standard Positioning Service, version 1.0, June 2014 (Indian Space Research Organization, Bangalore, 2014)
- 11.85 P. Misra, P. Enge: *Global Positioning System; Signals, Measurements and Performance*, 2nd edn. (Ganga-Jamuna Press, Lincoln, MA 2006)
- 11.86 A.S. Ganeshan: Overview of GNSS and Indian Navigation Program, GNSS User Meet. (ISRO Satellite Center, Bangalore 2012)
- 11.87 T. Neetha, A. Kartik, S.C. Ratnakar, A.S. Ganeshan: The IRNSS Navigation Message, J. Spacecr. Technol. **21**(1), 41–51 (2011)
- 11.88 O. Montenbruck, P. Steigenberger: The BeiDou Navigation Message, J. Glob. Position. Syst. **12**(1), 1–12 (2013)
- 11.89 T. Rethika, S. Mishra, S. Nirmala, S.C. Rathnakara, A.S. Ganeshan: Single frequency ionospheric error correction using coefficients generated from regional ionospheric data for IRNSS, Indian J. Radio Space Phys. **42**(3), 125–130 (2013)
- 11.90 H. Harde, M.R. Shahade, D. Badnore: Indian Regional Navigation System, Int. J. Res. Sci. Eng. **1**(SP1), 36–42 (2015)
- 11.91 T.S. Ganesh, C.K. Sharma, S. Venkateswarlu, G.J. Das, B.S. Chandrasekhar, S.K. Shivakumars: Use of two-way CDMA ranging for precise orbit determination of IRNSS satellites, Int. J. Syst. Technol. **3**(1), 127–137 (2010)
- 11.92 R. Babu, P. Mula, S.C. Ratnakara, A.S. Ganeshan: IRNSS satellite parameter estimation using com-

- 11.93 bination strategy, *Glob. J. Sci. Front. Res.* **15**(3), 1–10 (2015)
- 11.94 S. Kavitha, P. Mula, R. Babu, S.C. Ratnakara, A.S. Ganeshan: Adaptive extended Kalman filter for orbit estimation of GEO satellites, *J. Env. Earth Sci.* **5**(3), 1–10 (2015)
- 11.95 O. Montenbruck, P. Steigenberger: IRNSS orbit determination and broadcast ephemeris assessment, ION ITM, Dana Point (ION, Virginia 2015) pp. 185–193
- 11.96 PSLV-C22/IRNSS-1A brochure (ISRO, Bangalore 2013)
- 11.97 A. Kumari, K. Samal, D. Rajarajan, U. Swami, A. Kartik, R. Babu, S.C. Rathnakara, A.S. Ganeshan: Precise modeling of solar radiation pressure for IRNSS satellite, *J. Nat. Sci. Res.* **5**(3), 35–43 (2015)
- 11.98 K. Varma, D. Rajarajan, N. Tirmal, S.C. Rathnakara, A.S. Ganeshan: Modeling of IRNSS System Time-Offset with Respect to other GNSS, *Contr. Theory Inform.* **5**(2), 10–17 (2015)
- 11.99 N. Neelakantan: Overview of the Timing system planned for IRNSS, 5th Meet. Int. Comm. GNSS (ICG), Turn (UNOOSA, Vienna 2010) pp. 1–6
- 11.100 R.B. Langley: Dilution of precision, *GPS World* **10**(5), 52–59 (1999)
- 11.101 A.D. Sarma, Q. Sultana, V.S. Srinivas: Augmentation of Indian Regional Navigation Satellite System to improve dilution of precision, *J. Navig.* **63**(2), 313–321 (2010)
- 11.102 A.S. Ganeshan, S.C. Ratnakara, N. Srinivasan, B. Rajaram, K.N. Anbalagan: Tirmal: First position fix with IRNSS – Successful proof-of-concept demonstration, *Inside GNSS* **10**(4), 48–52 (2015)
- 11.103 N. Nadarajah, A. Khodabandeh, P.J.G. Teunissen: Assessing the IRNSS L5-signal in combination with GPS, Galileo, and QZSS L5/E5a-signals for positioning and navigation, *GPS Solutions* (2015), doi:[10.1007/s10291-015-0450-8](https://doi.org/10.1007/s10291-015-0450-8)

12. Satellite Based Augmentation Systems

Todd Walter

Satellite-based augmentation systems (SBASs) are designed to enhance the performance of standard global navigation satellite system (GNSS) positioning. SBASs improve the positioning accuracy by providing corrections for the largest error sources. More importantly, SBASs provide assured confidence bounds on these corrections that allows users to place integrity limits on their position errors. Several systems have been implemented around the world and several more are in development. They have been put into place by civil aviation authorities for the express purpose of enhancing air navigation services. However, SBAS services have been widely adopted by other user communities, as the signals are free of charge and easily integrated into GNSS receivers.

This chapter describes the basic architecture, functions, and application of SBAS. Because the key motivation behind SBAS is integrity, it is essential first to understand the error sources that affect GNSS and how they may vary with time or location. It is then explained how the corrections and confidence intervals are determined and applied by the user. The different SBASs that have been developed around the world are described and how they are developed to the same international standards such that each is interoperable with the others. The performances and services of each system are described. Finally, the evolution of SBAS from its current single-frequency single-constellation form into systems that support multiple-frequencies and multiple-constellations is described.

The goal of this chapter is to explain the motivation for developing SBASs and provide the reader with a working knowledge of how they function and how they may be used to enhance GNSS positioning accuracy and integrity.

12.1	Aircraft Guidance	340
12.1.1	Aviation Requirements	340
12.1.2	Traditional Navigational Aids	341
12.1.3	Receiver Autonomous Integrity Monitoring (RAIM)	341
12.1.4	Satellite-Based Augmentation Systems (SBAS)	342
12.2	GPS Error Sources	343
12.2.1	Satellite Clock and Ephemeris	344
12.2.2	Ionosphere	344
12.2.3	Troposphere	344
12.2.4	Multipath	345
12.2.5	Other Error Sources	345
12.3	SBAS Architecture	345
12.3.1	Reference Stations	345
12.3.2	Master Stations	346
12.3.3	Ground Uplink Stations and Geostationary Satellites	347
12.3.4	Operational Control Centers	349
12.4	SBAS Integrity	349
12.4.1	Integrity Certification	349
12.4.2	Threat Models	350
12.4.3	Overbounding	350
12.5	SBAS User Algorithms	351
12.5.1	Message Structure	351
12.5.2	Message Application	352
12.5.3	Protection Levels	353
12.6	Operational and Planned SBAS Systems	353
12.6.1	Wide Area Augmentation System (WAAS)	353
12.6.2	Multifunction Satellite Augmentation System (MSAS)	356
12.6.3	European Geostationary Navigation Overlay Service (EGNOS)	356
12.6.4	GPS Aided GEO Augmented Navigation (GAGAN)	356
12.6.5	System of Differential Corrections and Monitoring (SDCM)	358
12.6.6	BeiDou Satellite-Based Augmentation System (BDSBAS)	358
12.6.7	Korean Augmentation Satellite System (KASS)	358
12.7	Evolution of SBAS	358
12.7.1	Multiple Frequencies	358
12.7.2	Multiple Constellations	359
	References	360

12.1 Aircraft Guidance

Satellite navigation is finding ever increasing use in aviation. The utility of satellite navigation is enabled by a variety of augmentation systems. These augmentation systems are independent of the individual satellite constellations and monitor their performance continuously. Most importantly, the augmentations detect faults in real time and warn the pilots within seconds. Such assistance is needed because the constellation ground control system may not detect and report faults for tens of minutes or longer. The fault detection alternatives include aircraft-based augmentation systems (ABASs), ground-based augmentation systems (GBASs), and satellite-based augmentation systems (SBASs). This chapter focuses on SBASs [12.1]. Particular emphasis will be placed on the wide area augmentation system (WAAS, [12.2]), which is the SBAS for North America and was also the first operational SBAS. SBASs have also been developed in Japan, Europe, and India and are being developed in Russia, China, and South Korea.

Currently, SBASs augment the global positioning system (GPS) with the following three services:

- Integrity monitoring to improve safety
- A ranging function to improve availability and continuity
- Differential GPS corrections to improve accuracy.

Thus augmented, GPS meets the performance requirements for most phases of flight, including vertical guidance during airport approach. The first SBAS, WAAS, was commissioned in July 2003. The accuracy of the system rapidly made it an industry standard in GPS receivers. It routinely achieves horizontal accuracies better than 85 cm and vertical accuracies better than 1.2 m 95% of the time [12.3]. SBAS has achieved widespread adoption in nonaviation fields due to its open standard, free provision, and high accuracy [12.4, 5].

12.1.1 Aviation Requirements

Navigation systems used for aviation are judged by four key measures [12.1]:

1. Accuracy: The reported aircraft position must be close to the true position. Accuracy generally characterizes nominal errors and is usually expressed as a 95% confidence number. That is, it is specified as a number $\geq 95\%$ of the nominal position errors. Accuracy is the easiest requirement for SBAS to meet. Integrity, continuity, and availability are all much more difficult to achieve.

2. Integrity: An aviation navigation system must ensure that no position error larger than a maximum tolerable bound is presented to the pilot. All faults that could lead to larger position errors must be flagged within a specified time-to-alert (TTA), and the probability of failing to flag such a fault must be below some small probability per operation, typically between 10^{-5} and 10^{-9} depending on the operation.
3. Continuity: Once an aircraft begins a critical operation, the navigation system must continue to function until the operation is complete. The allowable probability of a navigation system outage during an aircraft approach operation varies from 10^{-5} to 10^{-9} per operation.
4. Availability: The navigation system must be functional and meet the above requirements a large fraction of the time in order to be useful to aircraft. Indeed, aviation requires availabilities better than 99–99.999% of the time. It is both unsafe and uneconomical to send an airplane to an airport only to discover once there that landing guidance is unavailable.

The numerical value for each requirement depends on the aircraft operation and they become more demanding as the aircraft is brought closer to other aircraft or to the ground, e.g., approaching an airport and preparing to land. So-called precision approach operations require vertical position accuracies of a few meters. Airport approach operations have particularly tough requirements for accuracy and integrity. Thus, most SBAS development and characterization effort focuses on this application. Generally, as long as the SBAS can meet the approach requirements, it will also meet the requirements for the other phases of flight.

Vertically guided approach is based on a smooth glide path with a constant rate of descent. This glide path, typically 3° , passes through a *decision height* where the pilot must decide whether or not to complete the landing. Pilots prefer vertically guided approach to the more challenging nonprecision approach. Nonprecision approach is also known as a step-down approach because pilots alternate a sequence of constant altitude segments with vertical step-downs as they approach the airport. This process requires the pilot to change the vertical descent rate of the aircraft at different points along the approach. This increased workload has contributed to a larger number of aircraft accidents. Before augmented GPS, an instrument landing system (ILS) or microwave landing system (MLS), sited at the airport, were the only systems able to provide precision

approach. SBAS enables precision approach without any airport-specific equipment. It allows the pilot to use a constant rate of descent down to a decision height of 200 ft above the ground, using a localizer precision with vertical guidance (LPV) procedure [12.6]. This is an important benefit as thousands of smaller airports are not equipped with ILS or MLS.

12.1.2 Traditional Navigational Aids

Traditionally, aviation has relied on radio-navigation signals from ground-based transmitters to determine aircraft position (Chap. 30). These systems are still largely in place and in wide use as the aviation community is very risk averse and slow to change. Typically, aircraft uses the same set of avionics originally installed in the aircraft, without update, for more than 20 years. It can be very difficult and costly to transition away from the existing set of equipment. The main set of nav aids currently in use is:

- Distance measuring equipment (DME)
- Very high frequency (VHF) omni-range (VOR)
- The tactical air navigation system (TACAN)
- The ILS.

These are described in more detail below.

A DME consists of a fixed antenna and transmitter–receiver that responds to aircraft interrogations after a fixed delay. As the aircraft knows the amount of the fixed delay, it obtains a true range to the antenna by subtracting the delay from the time between interrogation and response and dividing by 2. Because it also knows the location of the antenna, the aircraft then knows that it is somewhere on a circle at a fixed distance from that location. By querying 2 DMEs or using other information, the aircraft can further refine its position. The typical ranging accuracy of a DME is on the order of hundreds of meters. Each DME can be received to ≈ 150 nmi. The United States (US) Federal Aviation Administration (FAA) maintains ≈ 1100 DMEs in the conterminous United States (CONUS) to ensure nearly complete coverage by multiple DMEs.

A very high frequency (VHF) omnidirectional range (VOR) sends out two signals: one is uniform in all directions and the other is highly directional. By measuring the time between receipt of these messages, the aircraft can obtain a directional angle from the VOR. The typical accuracy of a VOR is less than half a degree. Thus, a co-located VOR and DME can provide an absolute position good to several hundred meters although this uncertainty increases with distance from the nav aids. As with DMEs, there are on the order of 1100 VORs within CONUS.

The TACAN system is a military version of a combined VOR/DME system. However, the DME portion of the TACAN signal is available for use to civilians and in the United States, most DMEs are actually TACANs. A VORTAC is a combined VOR and TACAN that meets both military and civilian needs.

An ILS consists of two sets of antennas and transmitters, one to provide angular offsets from the runway centerline and the other to provide angular offsets from the desired vertical glide path. The first set, called the localizer, provides horizontal guidance and the second set, called the glideslope, provides vertical guidance. In order to provide guidance to a single runway end, both the localizer and glideslope equipment are required. To serve both ends of a single runway, separate glideslope and localizer installations are required. Depending on the level of calibration, an ILS can safely guide an aircraft to within 200 ft of the ground (Category I) or all the way down to a blind landing (Category III). There are ≈ 1300 ILSs in the United States.

Each of these terrestrial navigational aids requires owned or leased land to occupy, reliable power and communication, maintenance, and constant calibration. Each piece of equipment is flight inspected for accuracy as often as every 2 months. The installation and ongoing support costs to maintain thousands of terrestrial navigational aids are significant. The FAA investigated satellite-based methods for providing guidance in order to reduce the existing nav aid infrastructure and overall costs of maintenance.

12.1.3 Receiver Autonomous Integrity Monitoring (RAIM)

The first and most common use of GPS in aviation provides horizontal guidance by utilizing receiver autonomous integrity monitoring (RAIM [12.7]), which is a variety of ABAS, to detect faults. Such RAIM-capable receivers estimate aircraft position and then compute the measurement residual for each satellite. The residuals are the difference between the actual measurement and the expected value that corresponds to the estimated position without using that satellite. This check detects measurement faults, provided there are at least 5 satellites in view with good geometry. RAIM is further capable of isolating measurement faults provided there are at least six satellites in view with very good geometry (Chap. 24).

GPS-based RAIM is the most widely used form of satellite navigation by aviation to date. It only provides horizontal guidance, but does so without any expensive ground infrastructure. Its coverage is global and not subject to limited ground networks or loss of signal due to blockage by terrain. It is also generally far more ac-

curate than VOR/DME/TACAN, and does not require flight inspection to maintain calibration.

RAIM leverages the normally over-specified nature of the GPS position solution, but this method is very sensitive to the state of the GPS constellation. In poor geometries, RAIM quickly becomes unavailable. For this reason, RAIM receivers cannot be the primary navigation aid; they must supplement another navigation aid. In contrast, SBAS-enabled receivers may be the primary navigation system for nonprecision approach because the fault monitoring is done on the ground and communicated to the aircraft. SBAS can provide availability in much worse geometries than RAIM can support.

12.1.4 Satellite-Based Augmentation Systems (SBAS)

SBAS is capable of assuring better horizontal accuracy than RAIM as well as providing vertical guidance. Therefore, SBAS can safely bring the aircraft closer to the ground with fewer satellites in the sky and with worse observational geometry.

An SBAS utilizes a network of ground monitors to continuously observe the performance of the navigation satellites. As shown in Fig. 12.1, the reference stations send their measurements to master stations that determine differential corrections and corresponding confidence bounds. Each master station processes the measurements and transmits data to an uplink station. The uplink station relays this information to the end users via a geostationary Earth orbit (GEO) satellite. Each SBAS has multiple master stations, uplink stations, and GEOs so that it can reliably survive the

failure of any one component. SBAS augments the core constellations with the following three services:

1. *Differential corrections*: SBAS broadcasts differential corrections for each satellite tracked by the ground network. The SBAS also transmits corrections for the effects of ionospheric delay over its region of interest. By applying these corrections to their pseudorange measurements, the user equipment improves its position accuracy.
2. *Integrity monitoring*: SBAS also broadcasts error bounds for each monitored satellite and each ionospheric correction parameter. These error bounds are used to determine the maximum possible airborne position error that may remain after the differential corrections are applied. Error bounds are appreciably more difficult to generate than differential corrections because the probability that the position error bound fails to overbound the true error must be smaller than 10^{-7} per approach. In addition, this information must be updated within 6 s of any unsafe condition.
3. *Ranging*: The SBAS GEO signals are similar to GPS L1 coarse/acquisition (C/A) signals in design and so an SBAS-enabled receiver uses essentially the same hardware as a normal GPS receiver. In addition, the SBAS signals are synchronized to GPS so they can be used for ranging. The additional ranging measurements are added to the suite of GPS ranging signals to improve the time availability and continuity of the position fix.

Each master station generates a grid of corrections for the ionosphere over its coverage region. This grid is 5° by 5° in latitude and longitude between 60° S and

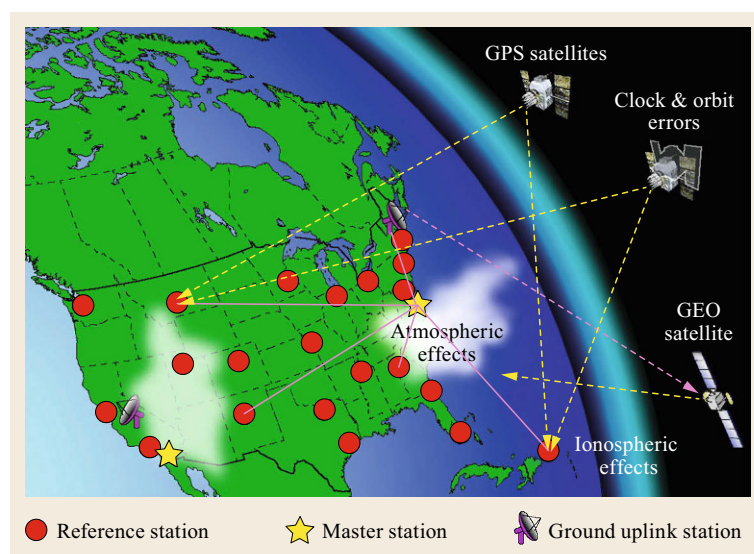


Fig. 12.1 General concept of a satellite-based augmentation system (SBAS)

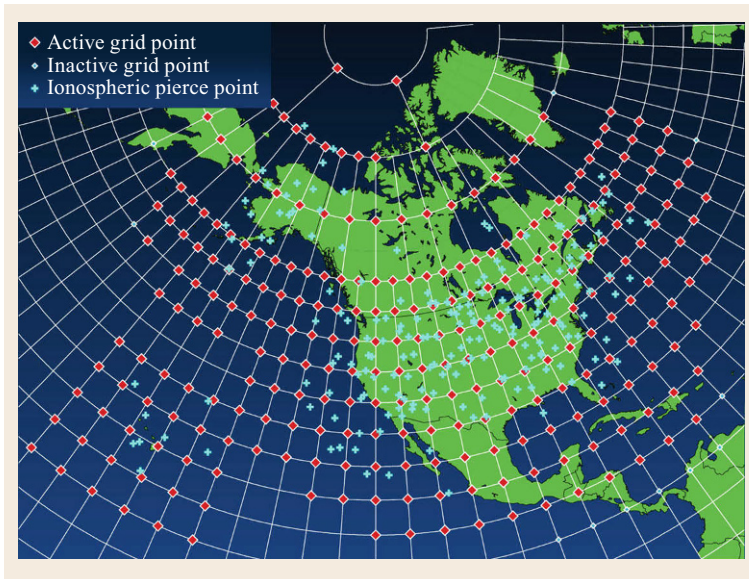


Fig. 12.2 SBAS ionospheric grid over North America

60° N and is less dense over the polar regions [12.8]. As with the stand-alone GPS single frequency ionosphere model, the SBAS ionospheric corrections model the ionosphere as though it were a thin shell existing at 350 km above the surface of the Earth [12.9]. The line of sight between the receiver and the satellite penetrates this layer at a point labeled the ionospheric pierce point (IPP). The user applies the four surrounding grid values to interpolate the ionospheric delay specific to each location of their IPPs. Figure 12.2 shows the SBAS ionospheric grid over North America with the grid points used by wide area augmentation systems (WAASs) indicated by *red diamonds*. Also shown are the IPPs as measured by the reference stations at one particular time.

The master station also generates a vector correction for each GPS satellite in view of the reference network. One element corrects the satellite clock and the other three elements are corrections for the three dimensions of satellite position. These corrections are generated from the pseudorange measurements after the ionospheric contribution has been removed and errors due to the troposphere and multipath have been minimized.

While the processing to generate the ionospheric and satellite specific corrections is sophisticated, the more difficult task is to bound the position errors that will remain after the corrections are applied. The bounds for the residual errors in the ionospheric corrections are called grid ionospheric vertical errors (GIVEs). The GIVE bounds the ionospheric correction for a given point in the grid for a line of sight that passes vertically through that point. Lines-of-sight at other angles get multiplied by a geometric obliquity factor to adjust the delay and confidence values for the longer ray path. The master station also bounds the impact of the satellite-specific error after correction, and these bounds are called user differential range errors (UDREs). They bound the projection of the satellite clock and location errors when projected onto the line-of-sight to the worst-case location in the coverage area.

The master station packs the ionospheric corrections, satellite specific corrections, and associated bounds into the SBAS message stream. This message stream is uplinked to the GEOs. These satellites are essentially bent pipes – they simply shift the uplink signal frequency and broadcast the message to users everywhere in the geostationary footprint.

12.2 GPS Error Sources

GPS signals are affected by many potential sources of error. It is important to understand these error sources and the possible effect they may have on the signal. For integrity, we are primarily interested in the effect they

have on ranging accuracy. The errors may cause unmodeled variations in the reception time of the signals and hence the apparent range to the satellite. An SBAS first attempts to correct these errors in order to improve

accuracy. Then, to the extent that it cannot fully correct the errors, it must describe to the user how much uncertainty remains on each of their corrected pseudorange measurements. It is therefore essential to understand and describe the physical source and effect of the different error sources. These sources are usually broken into three categories:

- Those originating on the satellite with the generation and broadcast of the signal.
- Those affecting the signal during its propagation from the satellite to the user.
- Those affecting the signal at the receiver and in its immediate vicinity.

The first category includes errors in the described satellite orbit position and clock offset, biases between the signals on different frequencies or between the code and carrier components, deformations of the signals, and look-angle-dependent biases from the satellite antenna. The propagation environment includes the effects of the ionosphere and the troposphere. The final category includes local multipath, receiver noise and tracking errors, and user antenna bias effects. For SBAS, both the errors affecting the reference stations and the users are important. The most significant error sources are described in greater detail below.

12.2.1 Satellite Clock and Ephemeris

Satellites suffer from nominal ephemeris and clock errors (Sect. 3.3.4) even when there are no faults present. These are typically very small for GPS, usually less than a meter in projected error. Occasionally, the broadcast GPS clock and ephemeris information may contain significant errors relative to the true state of the satellite position and clock. Such faults may appear as jumps, ramps, or higher order errors in the GPS clock, ephemeris, or both. These faults may be created by changes in state of the satellite orbit or clock, or simply due to the broadcasting of erroneous information. For example, a clock fault may lead to a sudden change in the timing of the broadcast signal while the position description remains accurate. Another example is an unannounced maneuver where the orbit suddenly changes, but the clock remains accurate. Alternatively, the satellite state may not change, but the navigation data that is broadcast to the user is changed to contain incorrect information. Another possibility is that everything about the satellite is correct, but either the user or the reference station incorrectly decodes the ephemeris information. For GPS, the nominal clock and position errors create projected pseudorange errors that typically have a standard deviation better than 1 m. Faults are rare

on GPS, typically occurring no more than twice a year, but may lead to projected errors of several kilometers.

12.2.2 Ionosphere

The ionosphere (Chap. 6) is a complex three-dimensional (3-D) distribution of free electrons primarily distributed between 100 and 1000 km above the surface of Earth [12.9]. It is often modeled as a two-dimensional (2-D) structure occurring in a thin shell at a height of 350 km. The electron distribution varies over the course of the day with a maximum effect in the local afternoon when the Sun's radiation has created the largest number of free electrons, and a minimum effect at night when those same electrons have recombined with the positive ions. There are seasonal changes with the ionosphere as the Earth's magnetic field changes its orientation with respect to the Sun. The Sun also undergoes an ≈ 22 year cycle where it reverses its magnetic field. This leads to an 11 year cycle in the ionosphere with significantly more ionospheric delay and disturbances near the maximum of the cycle than nearer to the minimum.

For nonequatorial regions (roughly $> 25^\circ$ of latitude), the thin shell model is usually a very good model. The ionosphere is easily estimated and bounded over large distances by assuming a linear variation in delay in the east–west and north–south directions. However, periods of disturbance occasionally occur where simple confidence bounds fall significantly short of bounding the true error [12.10, 11]. Additionally, in equatorial regions of the world, the ionosphere often contains significant 3-D structure. Disturbances can occur over very short baselines causing them to be difficult to describe in the limited SBAS message structure. Variations > 20 m of vertical delay over a 50 km baseline have been observed, as have rates of change as large as four vertical meters of delay per minute.

12.2.3 Troposphere

Tropospheric errors (Chap. 6) are typically small compared to ionospheric errors or satellite faults. Historical observations were used to formulate a model and analyze deviations from the assumed model used by SBAS. The assumed model will not exactly match the local climatological conditions. There are unpredictable variations in barometric pressure, temperature, and moisture content. Each of these variations may produce up to a few decimeters of vertical delay error that can map to a few meters of error at very low elevation angles. Typically, the total vertical error is < 10 cm and therefore < 1 m at low elevation [12.12].

12.2.4 Multipath

Multipath (Chap. 15) depends upon the environment surrounding the antenna and on the satellite locations. In aircraft, a well-placed antenna may have a very clean environment and the motion of the aircraft usually causes multipath to vary quickly. Thanks to carrier smoothing, the overall aircraft multipath can be reduced to < 25 cm standard deviation for a narrow correlator receiver [12.13]. The reference stations however may be in much more cluttered environments and therefore can experience multipath errors of several meters. Because the reference station antennas are stationary, the period of the multipath can be 10 min or greater. For GPS, multipath also contains a periodic component that repeats over a sidereal day. Thus, severe multipath may be seen repeatedly for several days or longer. However, with two frequencies, the reference stations may use a very long time constant for carrier smoothing and also be able to achieve standard deviations < 25 cm after sufficient smoothing time.

12.2.5 Other Error Sources

The previous sections describe the four most significant error sources, but there are other error sources that usually are not a factor but that potentially could affect performance. One such example is signal distor-

tions on the GPS codes [12.14, 15]. Because the signals are not strictly identical, there will be differences in their measured arrival time that depend upon the correlator spacing and bandwidth of the observing receivers. Such biases would be transparent to a network of identically configured receivers, but could be noticeably different to a user receiver with a different design. There are nominal deformation biases that are always present and may be several decimeters large. There is also concern over possible fault modes that could lead to errors > 10 m.

Another postulated threat is that a satellite may fail to maintain the coherency between the broadcast code and carrier. This fault mode is one that occurs on the satellite and is unrelated to incoherence caused by the ionosphere. This threat causes either a step or a rate of change between the code and carrier broadcast from the satellite. This threat has been observed on the L5 signals of the new Block IIF GPS satellites.

Look-angle-dependent biases in the code and on the carrier phase on both L1 and L2 are present on GPS antennas [12.16]. These biases may be several tens of centimeters. They may result from intrinsic antenna design as well as manufacturing variation. They are known to be present on the satellite antennas, on the reference station antennas, and on the users antennas. A closely related error source is survey error on the reference station antennas which could lead to errors in estimating the satellite corrections.

12.3 SBAS Architecture

As previously mentioned, an SBAS consists of three elements:

- The reference network
- The central processing facility
- The GEOs.

The reference network collects the basic GPS data in real-time and forward it for further analysis. The central processing facility evaluates the data and generates corrections and decisions about integrity. This information is then broadcast to the users via GEOs. These elements and their function are described in the following subsections.

12.3.1 Reference Stations

Each reference station contains independent threads of reference equipment. Each thread consists of an antenna, a dual-frequency GPS receiver, an atomic clock, and redundant communication links. Figure 12.3 shows racks containing three threads of equipment for

a WAAS reference station. Redundant threads are included so that hardware faults may be readily detected. The reference receivers are dual frequency. Every sec-



Fig. 12.3 A WAAS reference station (reproduced with permission of the FAA satellite navigation team)

and they take pseudorange measurements and carrier-phase measurements at the GPS L1 and L2 frequencies. The L1 and L2 frequencies are 1575.42 and 1227.60 MHz, respectively. The atomic clock makes it easier to compare previous measurements to the current ones and to identify outliers. The raw measurements from the reference stations are sent to each master station along redundant communication lines to ensure that each measurement arrives with very high likelihood. Reference stations are spaced ≈ 200 km or more apart. They are often placed at facilities that can provide security, reliable power (with backup), and reliable communications. No processing is performed at the reference station locations. Instead the raw measurements are all sent to the central locations for processing. The reference station network should have enough redundancy such that the loss of any individual station will not limit availability of the overall service.

12.3.2 Master Stations

An SBAS master station has four main tasks:

- Collect the data
- Formulate the corrections
- Determine the confidence bounds
- Pack the information into messages for broadcast.

The master station will seek to obtain all of the raw GPS data from every thread of every reference station. However, in order to meet the TTA it cannot wait too long for this data to come in. At a certain time it has to move forward and make its determinations with the information that it has for the epoch in question. This process is repeated every second as the master station continuously has to decide what the next message to send should be.

Upon getting the data from the receivers, the master station first performs consistency checks to identify and isolate erroneous data. The data from the parallel threads at each reference station must agree with each other and with previous information. If it does not, the master station must determine which information is incorrect. If it cannot do so, then it must immediately warn the user that the prior correction data may be unsafe. Much more commonly, it is able to identify and remove bad measurements before they are used downstream. Next the data is fed into various filters and estimators. There is a satellite clock and orbit estimator that also estimates reference station clock offsets. There is an estimator for the biases between the L1 and L2 signals that are induced by hardware on the satellites and at the reference stations. Finally, there is an estimator for the ionospheric delay at each of the grid points that the SBAS chooses to correct.

Most importantly, there are safety monitors that determine how much error may be present in the estimates. By correcting GPS with SBAS, one is initially doubling the risk of failure. There are now two very complex systems that can fail instead of just one. The first task for SBAS is self-monitoring to ensure that it does not introduce error. Each reference station contains parallel threads of equipment. Each thread operates independently of the others. As a first layer of screening for errors, the output of each thread is compared against the others. The expected geometry difference is first removed using the surveyed antenna coordinates and the broadcast satellite position. Additionally, the clock difference between each thread must be resolved from measurements. This is performed by combining information from all common satellites over time. If the corrected measurements disagree by too much then they are discarded. If too many measurements are discarded from a particular thread then it is flagged for maintenance. This cross comparison will identify any large receiver/clock failures as well as large multipath errors not common between the antennas. Smaller receiver errors or any common mode error can still escape detection. Later monitors compare the measurements from different reference stations for consistency as well as examine the temporal behavior to try to identify SBAS errors and prevent them from affecting performance. By screening measurements across multiple threads at the first stage, the vast majority of harmful errors are eliminated before they can affect downstream filters.

The monitors continue by characterizing the levels of code noise and multipath remaining on the measurements after error screening and carrier smoothing [12.13]. These screened measurements are used to monitor errors on the satellites and estimated delays due to ionosphere so it is very important to understand and bound the limits of observability. The confidence values associated with each measurement are then propagated through the subsequent monitors so that the monitors may accurately state how much certainty they have in their ability to screen for errors.

The postcorrection satellite clock and ephemeris errors are bounded by the UDREs. The master station looks at the projected clock and ephemeris error throughout the service volume and has to make sure that the UDRE is sufficiently large to protect all users. However, there are other errors that may be present on the satellite. The code and carrier signals may not be completely synchronized. Should this be the case, the measured range to the satellite will vary with the amount of smoothing that has been performed which can change for users depending on time of acquisition and most recent cycle slip. There is also the possibility of variations in the signals shape that can affect the

tracking of the signal. All of these possible error sources have to be covered by the UDRE.

Dual frequency measurements are required to generate the ionospheric corrections. The ionosphere is dispersive and so the ionospheric delay at L1 is different from the delay at L2. More specifically, the observed delay is inversely proportional to the frequency squared. The SBAS ground system leverages this relationship to estimate the ionospheric delay at the vertices in the grid. Unfortunately, the avionics directly cannot make use of the L2 signal because it lies in a nonaviation portion of the radio spectrum. The FAA cannot assure its availability. Hence, the ground system estimates the ionospheric delay for the avionics and sends the grid of ionospheric delay estimates to the airborne user. The density of the reference network is determined by the spatial decorrelation of the ionospheric delay [12.17]. Few reference stations are required if the ionosphere is always smooth. If the ionosphere has steep gradients, then a greater number is required. In the future, GNSS satellites will broadcast two signals for civil aviation (L1 and L5). At that time, new avionics will use both frequencies to compute the ionospheric delay in the aircraft because the L5 frequency does fall within a protected aviation band.

The ionospheric delay value at each grid point must be estimated from the individual ionospheric measurements from each reference station [12.18–21]. In addition, the ionospheric correction error at each grid point is bounded by the GIVE. Because the measurements do not coincide with the grid points they must be combined in a way to account for the potential spatial variation of the ionosphere. The GIVEs must account for the measurement errors and the uncertainty in the propagation model. The ionospheric grid points (IGPs) themselves are separated by roughly ≈ 500 km. Therefore, it is not possible to resolve very fine scale structure of the ionosphere. Further the measurements from the reference stations are sparse. The SBAS method for correcting the ionosphere depends on the fact that the ionosphere is generally slowly varying over hundreds of kilometers. The IGP delay algorithm nominally assumes limited variation in latitude and longitude, but must be prepared to identify times when its assumptions are invalid. Sometimes the ionosphere may be in a more disturbed state where the basic SBAS model is not an accurate description. At these times, the algorithm must recognize the problem and increase the confidence bounds accordingly [12.22].

SBAS uses a standard tropospheric model to predict the amount of tropospheric delay that exists on both the reference station and user lines of sight [12.12]. This is a climatological model based on years of primarily North American observations but that has been verified

after the fact with data from other parts of the world. It provides values for the barometric pressure, temperature, and other parameters given a latitude and time of year. From these parameters the amount of tropospheric delay can be estimated. This model also provides an upper bound on the error that may be remaining after applying it.

The satellite, ionospheric, and tropospheric corrections can be applied to each reference station measurement to evaluate the combined effect of the corrections for that specific line-of-sight. These errors should combine together in the expected manner. If the total error bound does not appear to properly bound all such measurements then the UDREs and GIVEs may need to be increased. This range domain check is another reasonability test to ensure all of the information is consistent.

As a final check, each reference station thread can evaluate its corrected position solution against the known surveyed location of its antenna. This helps to ensure that all of the corrections are working well together and are adequately bounded. The range domain and position domain tests ensure that all of the corrections combine together correctly. The integrity bounding methodology requires that the errors can be treated as though they are independent. A dependency that leads to a magnification of the errors in the position domain would become more obvious with these evaluations.

Finally, the message processor determines which 250 bit message should be sent for the current epoch and packages it appropriately [12.8, 23]. Usually the messages can follow an expected schedule. However, in the event of an integrity alert, the master station must send a message capable of alerting all affected users. If only a single satellite or IGP is affected, then the messages specific to those confidence bounds need to be sent. However, if many satellites are affected or the SBAS cannot autonomously isolate the faulty data, then more of the SBAS service may need to be alerted as potentially unsafe. Fortunately, such events are exceedingly rare. When alerts are broadcast, they are repeated four times in a row. This is due to the concern that a user receiver may miss an individual message (or even up to three messages). It is important to ensure that the user receives this data when prior information is no longer correct.

12.3.3 Ground Uplink Stations and Geostationary Satellites

Geostationary satellites are an excellent means for disseminating the SBAS messages. An SBAS can cover a large continental-scale region, as does the footprint of a GEO. The GEO signals are made very similar to

the GPS L1 C/A and L5 signals, respectively, on those frequencies. Thus, they provide extra ranging measurements for the user. These signals broadcast data at 250 bit/s, which is sufficient to transmit the SBAS corrections and confidences. The signals come from space and are therefore unlikely to be blocked by terrain in open sky environments where aircraft typically operate. The very name for SBAS, *satellite*-based augmentation system, comes from the pairing of the ground system with this satellite-based method of delivery. Figure 12.4 depicts the ANIK-F1R GEO used by WAAS.

The GEOs in use today are simple transponders. They listen for an analog signal at one frequency, translate it to the correct L-band frequency, and retransmit it toward Earth with minimum latency. The pseudo-random noise (PRN) code, messages, and timing are all generated on the ground. The signals are controlled through a closed-loop system that makes it appear as though they originated on the spacecraft [12.24]. The satellite effectively redirects the signal from the ground uplink station (GUS) back down toward the ground. The only change made by the satellite is from the uplink frequency to the correct downlink frequency. This approach is used because the transponder payloads are lighter and less expensive than the full navigation payloads on GPS satellites.

The GUS consists of a computer to receive messages from the multiple master stations, an atomic clock to provide a stable frequency reference, a signal generator to create the signal to uplink to the GEO, a receiver to monitor the GEO downlink signal, a GPS receiver to ensure the GEO is synchronized to GPS time, and a controller to steer the uplinked signal. Figure 12.5 shows the large antenna at the GUS in Napa valley, California, used by WAAS to uplink the signal to its GEO at 133° W. The ground uplink signal is most commonly > 3 GHz. The computer must decide which

message to send the next epoch. This will be based upon which master stations it has received messages from and which ones it has sent in the past. Typically, it will continuously send messages from the same master station. However, if communication to that station is interrupted, or if it is commanded to switch to another master station, then it will switch. If it receives no valid messages then it can either send an empty message or initiate an alert sequence.

The generated signal is very similar to a GPS L1 C/A code signal. The main differences are that the center frequency is well above the L1 band and the data bits are switched at 500 sps (symbols per second). The message is encoded onto the signal and it is beamed up to the GEO. The GEO receives this signal and down-converts it to L1 and broadcasts it back down to Earth. The signal is received at the GUS and the center frequency and timing of the chips on the uplink signal are adjusted to make it appear as though the downlink signal was generated on the GEO in synchronization with the GPS satellites.

The GEO signals are generally less accurate than the GPS signals. The transponders for some GEOs have a narrower bandwidth. This difference creates a loss of precision and some signal distortion. By generating

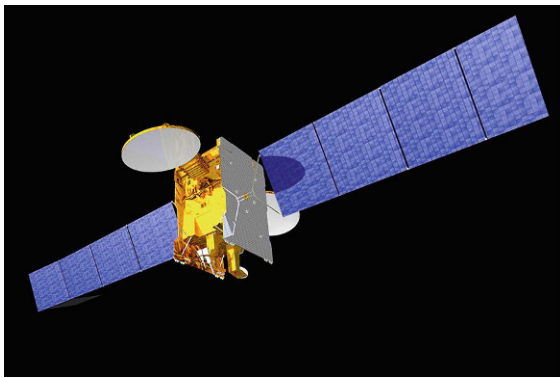


Fig. 12.4 The ANIK-F1R GEO used by WAAS (reproduced with permission of the the FAA satellite navigation team)



Fig. 12.5 A WAAS geostationary uplink station (reproduced with permission of the FAA satellite navigation team)

the signal on the ground some of the uplink path errors (e.g., ionosphere, troposphere) cannot be fully removed and therefore affect the downlink accuracy. Further, because the GEOs move very slowly in the sky, carrier smoothing does not reduce the multipath error on the ground at a static location, such as the reference receivers, very effectively. This increased error leads to less accurate orbit and clock estimation and larger uncertainty in bounding the error. Aircraft motion does cause enough variation for carrier smoothing to be effective in the aircraft.

12.3.4 Operational Control Centers

Figure 12.6 shows one of the two operational control centers for WAAS that has three master stations, one of which is located at this center. From this center, the operators can monitor the status and performance of WAAS. The operators schedule maintenance and upgrades of the various components at the reference, master, and uplink stations. This control center also monitors weather, air traffic, and the traditional navigational aids. The operators interact with other systems in the national airspace to ensure all

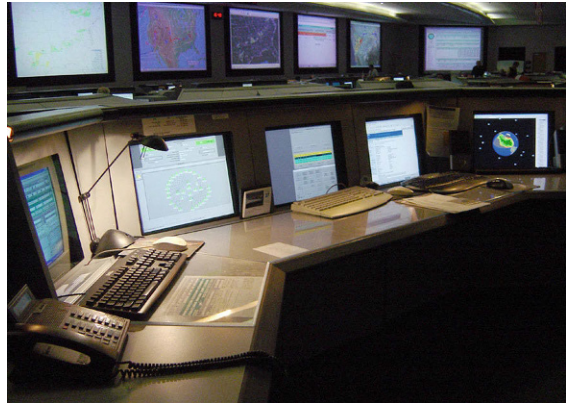


Fig. 12.6 A WAAS operational control center (reproduced with permission of the FAA satellite navigation team)

are well integrated. The different systems are managed together to ensure that any routine maintenance can be optimally scheduled and unplanned disruptions are properly communicated. The control center also produces notices to inform users of changes to the system performance [12.25] and interact with operators of GPS.

12.4 SBAS Integrity

Augmentation systems for aviation are very different from conventional differential GPS services. They are supplementing and ultimately replacing existing navigational aids whose safety has been demonstrated over many years of operational experience. Consequently, the safety of an augmentation system must be proven before it is put into service.

The integrity requirement is that the positioning error must be no greater than the positioning confidence bound, known as the protection level. This requirement is specified with a TTA and with a probability. The TTA requirement means that if the position error exceeds the protection level, the user must be informed within a very short period (6 s for the most demanding SBAS operation). Once a fault has occurred, the position error must fall below the protection level or the pilot must be informed that the system is unsafe to use within 6 s. The probability requirement is that no more than one in ten million operations may suffer an unannounced position error exceeding the protection level for > 6 s. SBAS provides differential corrections and confidence bounds to the user. The correction confidence bounds are used, together with the geometry of satellites tracked by the user, to calculate the protection level. In order to use the calculated position for nav-

igation, the protection level must be small enough to support the operation. The user only has real-time access to the protection level and does not know the true position error. Integrity is not maintained if the user has been told that the error in position is small enough to support the operation, but in fact, it is not. The majority of the effort in establishing an SBAS is in ensuring and demonstrating that these integrity requirements are met.

12.4.1 Integrity Certification

Certification is the process by which a provider ensures that the service it is providing meets the requirements. Certification involves analysis, testing, and documentation. Some of the important aspects of aviation integrity certification are [12.26]:

1. The aviation integrity requirement of 10^{-7} per approach applies to each and every operation. It is not an average over all conditions. The probability also applies to the worst allowed conditions.
2. Validated threat models are essential both to describe what the system protects against and to quantitatively assess how effectively it provides such protection.

3. The system design must be shown to be safe against all fault modes and external threats, including the potential for latent faults just beneath the system's ability to detect.
4. The small numbers associated with integrity analysis are not intuitive. Careful analysis must take priority over anecdotal evidence.

Because the requirement applies to all operations, threats and error conditions must be evaluated under worst allowable conditions. For example, if a user is allowed to operate under the maximum of the 11 year solar cycle, ionospheric errors must be modeled under this worst-case time. They cannot be an average of the high and low points of the solar cycle. Threat models are the means to capturing and describing the various errors that can occur and will be described in the next section. These models must describe both observed and anticipated threats and must treat them quantitatively.

12.4.2 Threat Models

Threat models describe the anticipated events against which the system must protect the user. The threat model must describe the specific nature of the threat, its magnitude, and its likelihood. Together, the various threat models must be comprehensive in describing all reasonable conditions under which the system might have difficulty protecting the user. Ultimately the threat models form a major part of the basis for determining if the system design meets its integrity requirement. Each individual threat must be fully mitigated to within its allocation. Only when it can be shown that each threat has been sufficiently addressed can the system be deemed safe. Quantitative assessment as opposed to a qualitative assessment is essential to establishing 10^{-7} integrity. SBAS works by analyzing specific failure modes and identifying which may be present and to what likelihood. Each potential failure mode must be ruled out within the limits of the system observability. If a failure mode is positively identified as being present in the system, or cannot be eliminated due to measurement noise, then the user must be notified within 6 s of it adversely affecting their position estimate.

SBAS is primarily thought of as addressing existing threats to GPS. However, it runs the risk of introducing threats in the absence of any GPS fault. By necessity, it is a complex system of hardware and software. Included in any threat model must be self-induced errors. Some of these errors are universal to any design while others are specific to the implementation. For example, the software design assurance of WAAS reference receivers is such that they cannot be trusted to be mistake free. Reference receiver software faults became

a unique threat that had to be mitigated through downstream integrity monitoring.

12.4.3 Overbounding

Each individual error source has some probability distribution associated with it. This distribution describes the likelihood of encountering a certain error value. Ideally, smaller errors are more likely than larger errors. Generally, this is true for most error sources. The central region of most error sources can be well described by a Gaussian distribution. That is, most errors are clustered about a mean (usually near zero) and the likelihood of being farther away from the mean falls off according to the well-known Gaussian model. This is often a consequence of the central-limit theorem, which states that distributions tend to approach Gaussian as more independent random variables are combined.

Unfortunately, the tails of observed error distributions rarely look Gaussian. Two competing effects tend to modify their behavior. The first is clipping. Because there are many cross-comparisons and reasonability checks within SBAS, the larger errors tend to be removed. Thus, for a truly Gaussian process, outlier removal would lead to fewer large errors than would otherwise be expected. The second, and more dominant effect is mixing. The error sources are rarely stationary. Thus, some of the time the error might be Gaussian with a certain mean and sigma and at other times it will have a different distribution. Because we do not necessarily have the ability to identify which condition is present at which time, different conditions will be aggregated into a single distribution. Such mixing may result from a change in the nominal conditions or from the introduction of a fault mode. Mixing generally leads to broader tails or large errors being more likely than otherwise expected.

Mixing causes additional problems. If the error processes were stationary, it would be possible to collect as large a data set as practical and then conservatively extrapolate the tail behavior using a Gaussian or other model. However, because the distribution changes over time, it is more difficult to predict the future performance based on the past behavior. Furthermore, mixing leads to more complicated distributions whose tails are more difficult to extrapolate.

Overbounding is the concept that an actual distribution can be conservatively described by a simple, usually Gaussian, model [12.27]. The overbounding distribution predicts that large errors are at least as likely to occur as they are for the true distribution. Even though the true distribution may not be completely known, there needs to be a practical way to represent it for analysis. Usually this involves a fair

amount of conservatism. A true distribution that is made up of a mix of zero-mean Gaussian distributions, could be overbounded by its constituent with the largest standard deviation. Thus, a real distribution that has a sigma value ranging between 1 and 2 m will be represented as though it were always 2 m (or perhaps 2.5 m for added protection). Through various overbounding theorems, the overbound also describes

how to combine the error with other terms that have been overbounded [12.27–29]. That is, SBAS will individually overbound the error for each satellite and each IGP. The GIVEs and UDREs broadcast by the SBASs describe overbounds of the actual error distributions affecting the corrections. The overbounding theorems allow users to combine these values to overbound the position errors by calculating the protection levels.

12.5 SBAS User Algorithms

The SBAS Minimum Operational Performance Standards (MOPS) are an internationally agreed upon document [12.8] that describes the method by which an SBAS transmits its differential GPS corrections and integrity information to users. The information is transmitted in 250 bit messages. These messages must be decoded and interpreted every second. The corrections are distributed across several individual messages. The corrections for individual satellites must be combined with receiver measurements and other local information to form the navigation solution and protection levels. The user must reconstruct and apply all of this information correctly. The MOPS ensures that all SBAS service providers encode their information in a compatible manner. The aviation receivers then know what to expect and will work with each of the different SBASs.

12.5.1 Message Structure

The broadcast message structure has 500 sps. These contain forward error correction to significantly reduce the risk of lost or misidentified bits [12.30]. The symbols go through a decoding process to produce 250 bit/s messages. There are two symbols for every message bit. The messages come once per second and contain 212 bit of correction data. Eight additional bits are used for acquisition and synchronization, 6 more bits to identify the message type and the remaining 24 bit designated for parity to protect against the use of corrupted data. Complementary message types must be stored and connected to the other individual components to form a single correction and confidence bound per satellite.

The SBAS messaging system contains the following elements:

- *Satellite corrections:* The SBAS broadcasts fast-corrections for satellite clock errors that may vary quickly in time. A fast-correction message corrects up to 13 satellite clock offsets and is sent every 6 s.

Clock offset rates of change are obtained by differencing sequential fast correction offsets. The SBAS also sends corrections for slowly varying satellite location and clock errors. The corrections consist of Δx , Δy , and Δz satellite locations (and possibly velocities) plus delta clock (and possibly delta clock rate). These long-term corrections are sent approximately every 2 min. Each long-term correction message corrects either two or four satellites depending on whether the rate of change information is also included.

- *Ionospheric corrections:* The SBAS broadcasts a grid of ionospheric corrections. Each ionospheric correction message updates the vertical delay estimate at up to 15 ionospheric grid points and is broadcast once every 5 min.
- *Confidence bounds:* In addition to the corrections, confidence bounds on the remaining errors are also broadcast. The UDREs must be sent every 6 s while the GIVEs are only updated every 5 min. These bounds are essential to maintain the integrity and TTA of the system. The UDREs are included in the fast-correction messages and the GIVEs are included with the ionospheric correction messages. In addition, there are messages that can provide the full 4×4 covariance matrix information for the clock/ephemeris error for each satellite. If broadcast, these matrices are sent every 2 min in messages that update two satellites. The covariance matrices are sent in a message labeled as message type 28 and are often referred to as message type 28 (MT28) parameters [12.31]. An alternate message can define regions where the UDRE values are to be increased. This regional information is sent in message type 27 and labeled as MT27 parameters. A system will either use MT28 or MT27 but never both as they both serve similar purposes but via different means.
- *Degradation parameters:* The potential error in the corrections increases over time. Parameters are broadcast to model these effects. The users apply

these degradations as their corrections age. They are particularly important in maintaining integrity and availability when a user misses the most recent correction.

- **Masks:** The PRN mask is used to designate which satellite belongs to which slot in the fast-correction messages. A mask is used to assign slots so that satellite identifications need not be sent with every fast correction. A similar ionospheric mask is used to associate each slot in the ionospheric correction message with a geographic grid point location. The use of masks reduces the required throughput because the masks are sent infrequently. They also inform the user as to which satellites and which IGP are corrected by the specific SBAS.
- **Geostationary navigation message:** In contrast to the GPS satellites, the current SBAS satellites are geostationary. Consequently, their location (ephemeris) does not need to be updated as frequently. Nor do they require as large a dynamic range as the geostationary orbit is restricted to a limited region about the equator. These messages broadcast the absolute position (x , y , and z in an Earth centered Earth fixed (ECEF) reference frame), as well as the velocity and acceleration values. The absolute clock and clock rate are also included in this message, as well as the reference time, and an issue of data. This message is broadcast every 2 min.
- **Preamble:** In order to allow the receiver to find the start of the 250 bit message an 8 bit preamble is included at the beginning of every message. There are three unique preambles that repeat in a fixed sequence. By searching for this specific bit pattern, a receiver can synchronize itself with the SBAS data signal.
- **Parity:** For data integrity, the SBAS must use a much stronger error detection algorithm than the six parity bits used in the GPS navigation message. Nonetheless, the overhead for error detection is reduced because the parity bits apply to longer messages than for GPS. The 24 bit cyclic redundancy check (CRC) ensures that the message the user applies is the one intended. Any bits corrupted in transmission are detected before they can create erroneous information.
- **Forward error correction:** Forward error correction is used so that the SBAS can send significantly more data than the 50 bit/s carried in the GPS navigation message. The chosen code for SBAS is a rate 1/2 convolutional code with a constraint length of 7.

The currently used SBAS messages are listed in Table 12.1.

12.5.2 Message Application

The user requires one long-term and two fast corrections for each satellite that it uses [12.23]. The two fast-corrections are differenced to determine the rate of change of the fast clock term. This rate is used with the most recent fast-correction to determine the fast clock value for the current time. The fast clock correction is added to the long-term clock correction to obtain the full clock correction. The orbit corrections are also taken from the long-term correction and added to the orbital location broadcast in the navigation message received directly from the GPS satellite. The UDRE is taken from the fast-correction and increased depending on how long ago the fast correction was received. If MT28 is used, the parameters are used to determine a scaled covariance matrix. This matrix, together with the normalized four-dimensional (4-D) line-of-sight vector, is used to determine a multiplier for the UDRE. Generally, when the user is close to the reference station locations, this scaling factor will be smaller. If the user is far from the reference stations, then this factor may significantly increase its product

Table 12.1 SBAS message types

Type	Contents
0	Do not use for safety applications (WAAS testing)
1	PRN mask assignments, set up to 51 of 212 bit
2–5	Fast pseudorange corrections and UDREs
6	Integrity information, UDREs (multiple satellite alert)
7	Fast correction degradation factor
8	Reserved for future messages
9	GEO navigation message (X, Y, Z, time, etc.)
10	Degradation parameters
11	Reserved for future messages
12	WAAS network time/Coordinated Universal Time (UTC) offset parameters
13–16	Reserved for future messages
17	GEO satellite almanacs
18	Ionospheric grid point masks
19–23	Reserved for future messages
24	Mixed fast/long-term satellite corrections
25	Long-term satellite corrections
26	Ionospheric delay estimates and GIVEs
27	WAAS service message
28	Clock-epemeris covariance matrix message
29–61	Reserved for future messages
62	Internal test message
63	Null message

with the UDRE. If MT27 is used, then the UDRE will be multiplied by those parameters depending on the user's location. As with MT28, the UDRE is generally multiplied by a larger term if the user is farther from the reference stations. This product, which also will include degradation terms, is expressed as a one-sigma value and referred to as the fast and long-term correction bound or s_{flt} .

Satellite corrections and confidences are all that are required for less precise lateral navigation. The user can apply the simple single-frequency ionosphere model broadcast by GPS and determine horizontal positions bounded to within a fraction 1 nmi. However, to obtain precise vertical guidance, the user must also apply the SBAS ionospheric corrections.

For each of their IPPs, the user must identify the surrounding IGP and obtain ionospheric delays for each. It requires a minimum of three enclosing IGPs, but ideally the four defining a rectangle about the IPP are all available. The user then applies a bi-linear interpolation of the surrounding delay values to obtain the vertical delay estimate at the IPP. It applies the same interpolation to obtain the user ionospheric vertical error (UIVE) bound from the surrounding GIVEs. These are converted from vertical to slant by applying the obliquity factor. The resulting confidence term is now called the user ionospheric range error (UIRE). The delay value is subtracted from the pseudorange measurement to that satellite with the corresponding IPP. The user also subtracts the MOPS-specified tropospheric model delay estimate from each line-of-sight to fully correct the range error.

12.6 Operational and Planned SBAS Systems

Four SBASs have been implemented around the world and at least three more are under development. The operational systems are all compatible with the MOPS and with existing certified SBAS receivers, but they are not identical. They have been developed specifically for their own regions and sometimes faced unique challenges. However, despite any differences SBAS receivers will work equally well with any of these systems and should be able to seamlessly transition from one to any other.

12.6.1 Wide Area Augmentation System (WAAS)

The WAAS has been fully operational for safety-of-life services since July 2003 [12.2]. It consists of 20 WAAS reference stations (WRS) in the CONUS, in addition to seven in Alaska, one in Hawaii, one in Puerto Rico, four

12.5.3 Protection Levels

The basic notion of the protection level equations is that the error sources are approximately Gaussian and that a Gaussian model is sufficiently accurate to be able to conservatively describe the positioning errors. Four error terms are used to describe satellite clock and ephemeris errors (σ_{flt}), ionospheric delay errors (σ_{UIRE}), tropospheric delay errors (σ_{tropo}), and airborne receiver and multipath errors (σ_{air}). The conservative variances of these terms are combined to form a conservative variance for the individual pseudorange error.

$$\sigma_i^2 = \sigma_{flt,i}^2 + \sigma_{UIRE,i}^2 + \sigma_{tropo,i}^2 + \sigma_{air,i}^2 \quad (12.1)$$

This pseudorange variance is inverted and placed on the diagonal elements of the weighting matrix, \mathbf{W} , and combined with the geometry matrix, \mathbf{G} , to form the covariance of the position estimate.

$$(\mathbf{G}^T \mathbf{W} \mathbf{G})^{-1} \quad (12.2)$$

Here the geometry matrix is expressed in a local east, north, up reference frame. The third diagonal element represents the conservative estimate of the error variance in the vertical direction. Since the vertical protection level (VPL) is intended to bound 99.99999% of errors it is set to the equivalent Gaussian tail value of 5.33. Thus, the final VPL for L1-only SBAS is given by

$$\text{VPL} = 5.33 \sqrt{[(\mathbf{G}^T \mathbf{W} \mathbf{G})^{-1}]_{3,3}} \quad (12.3)$$

in Canada, and five in Mexico for a total of 38. There are 3 WAAS master stations (WMS) and 3 geostationary satellites (GEOs). The GEOs are the Intelsat Galaxy XV satellite at 133° W (PRN 135), the Telesat ANIK F1R satellite at 107° W (PRN 138), and the Inmarsat-4 F3 at 98° W (PRN 133). WAAS is also in the process of procuring replacement GEOs including the SATMEX-9 which will become active in 2017 and be located at 117° W (PRN 131). A full list of all SBAS GEOs can be found in Table 12.2.

Figure 12.7 shows the reference station networks for all of the current and some of the developing SBASs. As can be seen, there is good sampling around the northern hemisphere. WAAS provides excellent coverage for lateral navigation.

Figure 12.8 shows the lateral navigation coverage provided by WAAS, as well as the Japanese and European SBASs. It can be seen that all of North America

Table 12.2 SBAS GEOs (after [12.32], courtesy of the US Air Force)

PRN	SBAS	Satellite	Location
120	European Geostationary Navigation Overlay Service (EGNOS)	INMARSAT 3F2	15.5° W
121	EGNOS	INMARSAT 3F5	25° E
122	Unallocated		
123	EGNOS	ASTRA 5B	31.5° E
124	EGNOS	Reserved	
125	System for Differential Correction and Monitoring (SDCM)	Luch-5A	16° W
126	EGNOS	INMARSAT 4F2	25° E
127	GPS Aided GEO Augmented Navigation (GAGAN)	GSAT-8	55° E
128	GAGAN	GSAT-10	83° E
129	Multi-function Satellite Augmentation System (MSAS)	MTSAT-1R (or -2)	140° E
130	Unallocated		–
131	WAAS	Satmex 9	117° W
132	Unallocated		
133	WAAS	INMARSAT 4F3	98° W
134	Unallocated		
135	WAAS	Intelsat Galaxy XV	133° W
136	EGNOS	ASTRA 4B	5° E
137	MSAS	MTSAT-2 (or -1R)	145° E
138	WAAS	ANIK-F1R	107.3° W
139	GAGAN	GSAT-15	93.5° E
140	SDCM	Luch-5B	95° E
141	SDCM	Luch-4	167° E
142–158	Unallocated		

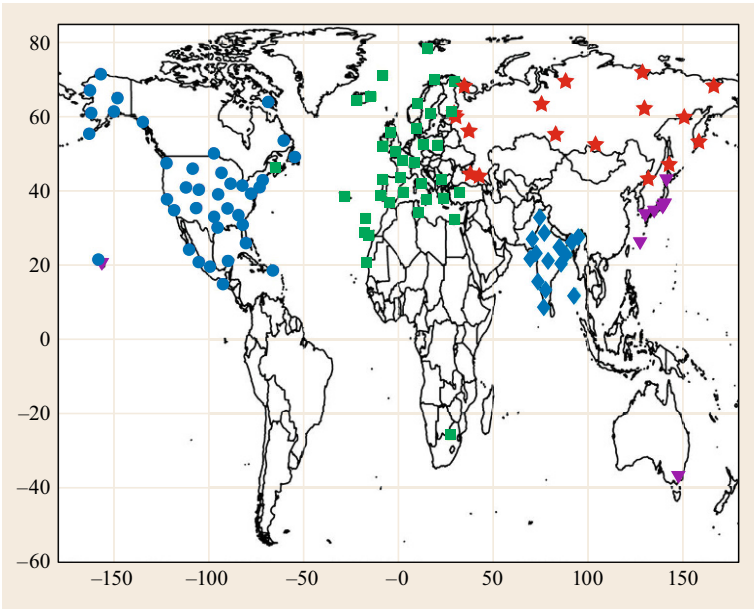


Fig. 12.7 Reference station networks of WAAS (dark blue circles), EGNOS (green squares), SDCM (red asterisks), and GAGAN (blue diamonds)

and part of South America can rely on WAAS to navigate to and from any airport of choice. The edges of GEO footprints can be seen in the north of this figure as visibility to at least one of the GEOs is a requirement to get SBAS service. Vertical guidance requires the precise ionospheric corrections of the grid and therefore is

restricted to a much tighter region around the reference stations.

Figure 12.9 shows the vertical guidance coverage area for the same three SBASs. It can be seen that WAAS covers CONUS, Alaska, and much of Canada and Mexico. Figure 12.10 shows the vertical coverage

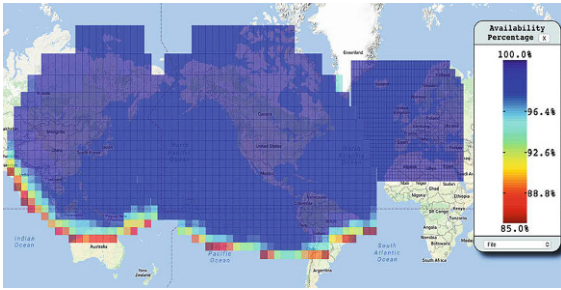


Fig. 12.8 Lateral navigation coverage for WAAS, MSAS, and EGNOS (after [12.33], reproduced with permission of the William J. Hughes FAA Technical Center)

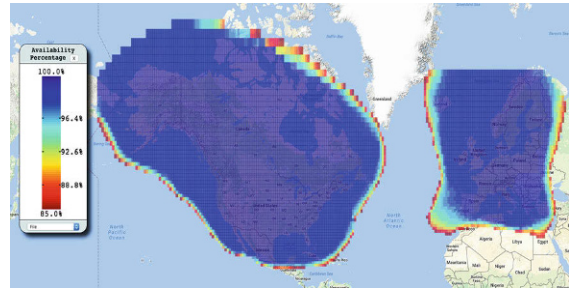


Fig. 12.9 Vertical navigation coverage for WAAS and EGNOS (after [12.33], reproduced with permission of the William J. Hughes FAA Technical Center)

area in greater detail and also tabulates the percentage of each region that achieves a certain level of availability. On the day shown in Fig. 12.10, 100% of CONUS has 100% availability, while 95.04% of Alaska has 99.9% or better availability.

Accuracy is very good when using WAAS. Horizontal accuracy is ≈ 0.75 m 95% in CONUS. This number can be compared to 3.2 m for uncorrected GPS under moderate ionospheric conditions. Under more severe ionospheric conditions the GPS positioning errors can be noticeably worse, but the WAAS corrected accuracy only worsens slightly. For example, in 2003 during a worse solar maximum period, uncorrected horizontal accuracy was 4.8 m 95% while WAAS corrected accuracy was 0.88 m. Horizontal accuracy in Alaska is also ≈ 0.75 m 95% but it is slightly worse in Mexico

(≈ 0.90 m) and Canada (≈ 1.0 m). Vertical accuracy for WAAS is ≈ 1.1 m 95% in CONUS compared to 7.6 m for uncorrected GPS. In Alaska, vertical accuracy is ≈ 1.3 m 95%. Again, somewhat worse 95% vertical performance is seen in Mexico (≈ 2.0 m) and Canada (≈ 1.5 m).

WAAS was fielded because its advantages relative to conventional nav aids were enormous. It has made precision vertical guidance available throughout the majority of North America. No local airport infrastructure is required for this service. Already more than 80 000 WAAS-enabled aviation receivers have been sold. More than 3500 vertically guided approaches have been commissioned. This is nearly three times as many as provided by ILS. WAAS allows users to access more than 2000 airports that had no previous instrument

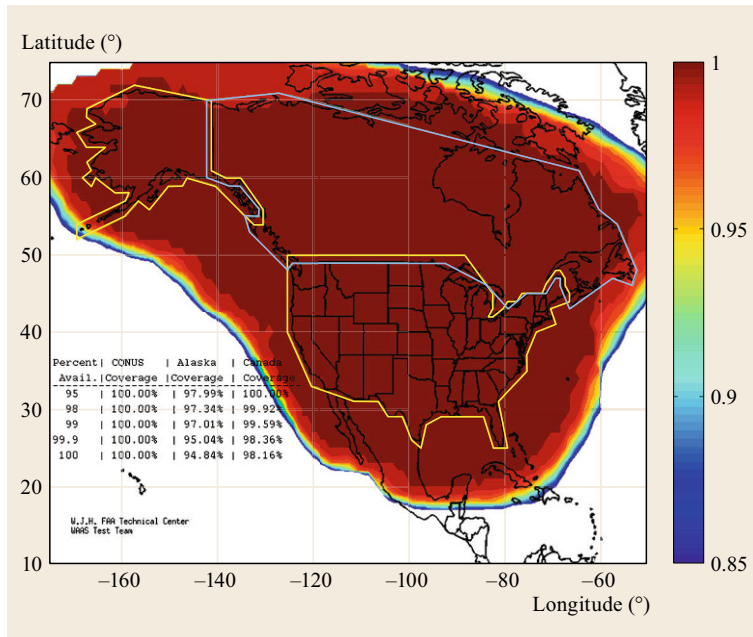


Fig. 12.10 Detailed vertical navigation coverage for WAAS on 19 March 2015 (after [12.33], reproduced with permission of the William J. Hughes FAA Technical Center)

approach. It is also widely used in nonaviation applications [12.34].

Agriculture uses SBAS to more accurately position vehicles and reduce fertilizer and pesticide use. Maritime uses SBAS to guide ships more accurately in poor visibility conditions. SBAS is incorporated into nearly every cell phone as the GEOs are widely visible, the correction data is freely provided, and the system accuracy is greatly improved.

12.6.2 Multifunction Satellite Augmentation System (MSAS)

The multifunction satellite augmentation system (MSAS) consists of six ground monitoring stations (GMSs) on the Japanese islands, in addition to one in Australia, and one in Hawaii for a total of eight [12.35, 36]. The station locations are shown as magenta triangles in Fig. 12.7. There are two master control stations (MCSs) and two multifunction transport satellite (MTSAT) GEOs at 140°E and 145°E. MSAS was commissioned for safety-of-life service in September 2007.

Due to the limited network size, the GEO UDREs for MSAS are set to 50 m and therefore do not benefit vertical guidance. Further the limited ionospheric observations offer little availability of vertical service. As a result vertically guided operations have not yet been authorized based upon MSAS. The Japanese civil aviation bureau (JCAB) has studied performance improvements that could allow it to provide vertically guided operations. Until then, MSAS provides only lateral guidance. Like WAAS, lateral guidance is available for quite a large region around and away from the reference stations.

12.6.3 European Geostationary Navigation Overlay Service (EGNOS)

The European geostationary navigation overlay service (EGNOS) consists of 39 ranging and integrity monitoring stations (RIMSs) in Europe, Africa, and North America [12.37–39]. The station locations are shown as green squares in Fig. 12.7. EGNOS has four master control centers (MCCs) and six navigation land Earth stations (NLESs) that control their three GEOs. The two operational GEOs are the Imarsat-3 F2 satellite at 15.5°W (PRN 120) and the Inmarsat-4 F2 satellite at 25°E (PRN 126). There is an additional GEO, Astra 4B at 5°E (PRN 136), that is under test and will become operational in 2015. Astra 5B is being procured for operation at 31.5°E (PRN 123) beginning in 2016. EGNOS was declared operational in Octo-

ber 2009, and was certified for safety-of-life service in March 2011.

EGNOS also has very good accuracy. It achieves 1.2 m 95% horizontal and 1.8 m 95% vertical accuracy over Europe. For a variety of reasons EGNOS has chosen to implement its GEO satellites without a ranging capability. They are only providing differential corrections and integrity information. It is possible that future GEOs will include ranging.

EGNOS currently implements message type 27 (MT27) rather than message type 28 (MT28) as used by WAAS, MSAS, and the SBAS in India. MT27 restricts the use of small UDRE values to a box centered on the European region (from 20°N to 70°N and 40°W to 40°E). The edges of this MT27 box can be clearly seen in Figs. 12.8 and 12.9. MT27 limits lateral navigation service more so than does MT28, but there is still excellent horizontal coverage in and around Europe. Availability of vertical guidance is very high for most of Europe as shown in Fig. 12.9. Figure 12.11 provides a more detailed view of the availability of vertical guidance over Europe. Over 175 vertical approaches have already been implemented that serve over 100 airports and hundreds more are under development.

EGNOS was developed with a greater emphasis on multimodal support [12.40]. The other SBASs were implemented by their local civil aviation authorities (CAAs) and are therefore originally designed to support aviation needs. They do in fact support all modes of transportation, however, aviation is their only mandate. EGNOS has a mandate also to support other modes of transport such as maritime, rail, and automotive. EGNOS was also originally designed to incorporate the Russian global navigation satellite system (GLONASS). Although GLONASS is not used for its safety-of-life service, GLONASS is still monitored and its measurements are available for other purposes. EGNOS makes its data and corrections available through its EGNOS data access service (EDAS). EDAS provides near real-time access to the RIMS measurements and the broadcast messages. Thus, EGNOS is available to users who do not have visibility to any of its GEOs [12.5, 41].

12.6.4 GPS Aided GEO Augmented Navigation (GAGAN)

India is developing the GPS-aided GEO augmented navigation (GAGAN) system [12.42, 43]. Currently it has 15 Indian reference stations (INRES) all in India. The station locations are shown as blue diamonds in Fig. 12.7. There are two Indian master control cen-

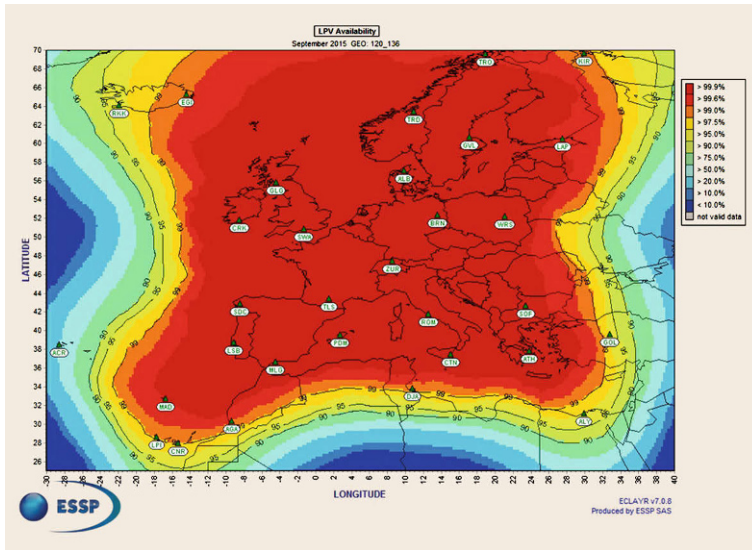


Fig. 12.11 Detailed vertical navigation coverage for EGNOS in September 2015 (after [12.39], reproduced with permission by the European satellite services provider (ESSP))

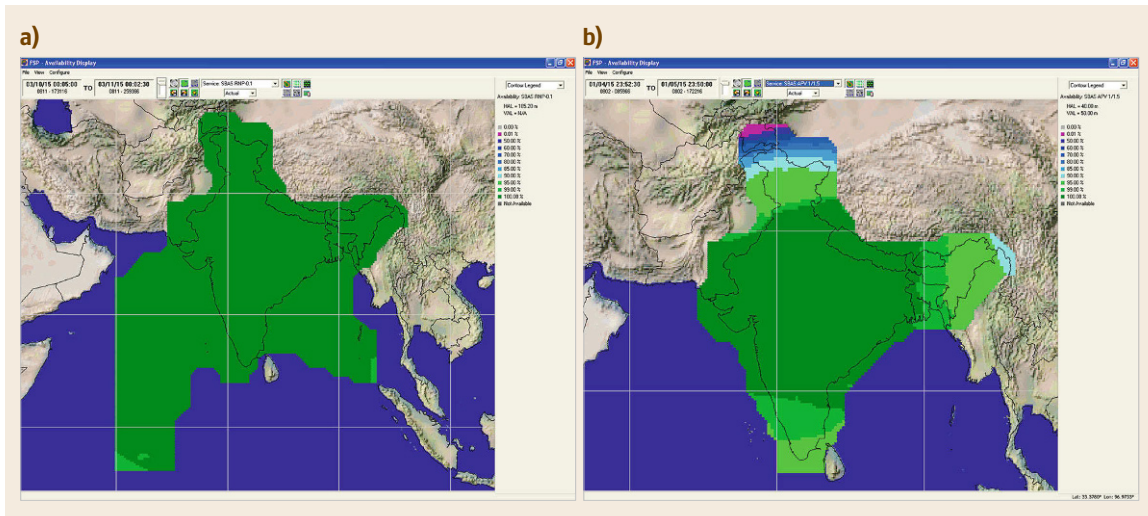


Fig. 12.12 (a) Lateral navigation coverage for GAGAN, (b) vertical navigation coverage for GAGAN (after [12.42], reproduced with permission by the Airports Authority of India)

ters (INMCCs), and three Indian navigation land uplink stations (INLUSs) to control its GEOs. GAGAN uses GSAT-8 at 55° E (PRN 128) and GSAT-10 at 83° E (PRN 127) as its GEOs. GSAT-15 at 93.5° E (PRN 139) is currently being deployed and will be launched in 2015.

The geomagnetic equator passes through India and GAGAN therefore faces the full impact of the equatorial ionosphere. During peak ionospheric activity, vertical guidance is not always available. Within the equatorial region, post-sunset hours are frequently beset by large depletions and scintillation. The depletions

create large gradients in the ionospheric delay that cannot be easily modeled by the SBAS thin-shell grid. Scintillations interrupt tracking to the satellite signals. Fortunately lateral navigation is less susceptible to these issues as service can be provided even with larger ionospheric delay uncertainty and fewer satellites in view. Figure 12.12a shows lateral availability within the Indian airspace.

Vertical guidance does require small ionospheric delay uncertainty and very good geometry. Thus, many evenings suffer a loss of service, especially during solar maximum periods. The current solar cycle reached its

maximum in 2014. Figure 12.12b shows vertical availability on a very good day. The advent of L5 will allow GAGAN to obtain high LPV-200 availability throughout all periods of the solar cycle. GAGAN was certified for lateral only service, which begun in February 2014. The vertical guidance service was certified in April of 2015.

The equatorial ionosphere also causes accuracy to be somewhat worse for GAGAN compared to the other SBASs located in mid-latitude. GAGAN achieves 2.3 m 95% horizontal and 3.7 m 95% vertical accuracy over India.

12.6.5 System of Differential Corrections and Monitoring (SDCM)

Russia is developing its system of differential corrections and monitoring (SDCM [12.44]). Currently SDCM has 19 prototype measuring points (MPs) in Russia and four prototype stations are available outside of Russia. The Russian station locations are shown as red stars in Fig. 12.7. There are also plans to use three GEOs: Luch-5A is at 16° W (PRN 125), Luch-5B at 95° E (PRN 140), and Luch-4 at 167° E (PRN 141). SDCM intends to add 27 additional MPs within Russia and three more outside of Russia. SDCM is still in its development phase with initial service expected in 2016 and fully certified service in 2019. SDCM further intends to augment both GPS and GLONASS. The SDCM prototype achieves 1 m 95% horizontal and 1.5 m 95% vertical accuracy over Russia.

12.7 Evolution of SBAS

Recently launched GPS satellites have two civil signals at protected aviation frequencies. When both signals are operational, they will allow users to measure the ionosphere directly instead of relying on the SBAS grid for corrections. The uncertainty of the users' direct ionospheric measurements will be much smaller than the broadcast SBAS confidence. Additionally, users can make these measurements anywhere, not just near reference stations. As a result, the service level will improve and the region of coverage will expand much farther from the SBAS networks. Further, additional constellations of navigation satellites are being fielded. The number of useful ranging sources for the user will soon dramatically increase. SBASs will be updated to take advantage of these improvements as they become available [12.47, 48]. The user will experience better availability for existing services and new, even

12.6.6 BeiDou Satellite-Based Augmentation System (BDSBAS)

China's satellite navigation system is called Beidou. It includes SBAS-like terms in its navigation message, but these are not backward compatible with existing SBAS receivers. The level of safety associated with these terms is not yet known. A mask and 18 parameters labeled as UDREs are included in one of its subframe messages. An ionospheric grid is also defined over China and delay values and parameters labeled GIVEs are included in other subframes. China has recently announced that it intends to also provide an SBAS service that is compatible with the International Civil Aviation Organization (ICAO) standard to be called BeiDou satellite-based augmentation system (BDSBAS [12.45]). It currently has 20 prototype ground stations in China with plans to ultimately have 30 stations within China and 20 more stations in surrounding areas. GEOs are planned for 80° E, 110° E, and 140° E. The ICAO compatible service is expected to be available around 2020.

12.6.7 Korean Augmentation Satellite System (KASS)

The Republic of Korea has also announced its intention to develop its own SBAS [12.46]. This SBAS will consist of five or more reference stations, two central processing facilities, four GUSs and two GEOs. KASS is in the early development stage with a plan to have preliminary service by 2020.

more demanding, services will likely become available.

12.7.1 Multiple Frequencies

The GPS satellites now being launched contain a new civil aviation signal. L5 is centered at 1176.45 MHz and is in a protected aviation band. As such, it will be approved for navigation when it becomes fully operational. When the L5 signal is used in combination with L1, the ionospheric delay for each line-of-sight can be directly estimated and removed. Removing the ionospheric delay will dramatically lower the uncertainty of the pseudorange measurement. Thus, if the SBAS is upgraded to provide satellite clock corrections appropriate for an L1/L5 user and the user similarly upgrades their avionics, SBAS service can be

dramatically expanded beyond the current grid of corrections [12.49].

Another important advantage of the second civil frequency is its relative immunity to ionospheric disturbances that are not well modeled by the MOPS grid. Because the user is now directly eliminating the amount of delay they actually experience, they are no longer affected by shortcomings in the MOPS ionospheric model. Thus, a dual frequency user would also have good availability in equatorial areas, even during peak solar activity. The weaker effect of scintillation may have some impact, however, we do not expect to lose vertical guidance altogether, at least not over large areas and for many hours [12.50]. Furthermore, the availability of two civil frequencies offers some protection against unintentional interference. If either L1 or L5 is jammed, the user still has access to guidance on the remaining frequency.

At the moment, the MOPS for an L1/L5 user are at the very early stage of development, so any ground or user improvements are still speculative. When a user has access to two civil frequencies, they can remove the ionospheric effects by forming the ionosphere-free combination of the two pseudoranges

$$p_{\text{iono_free}} = \frac{f_1^2 p_1 - f_5^2 p_5}{f_1^2 - f_5^2},$$

$$\sigma_{\text{iono_free}}^2 = \left(\frac{f_1^2}{f_1^2 - f_5^2} \right)^2 \sigma_1^2 + \left(\frac{f_5^2}{f_1^2 - f_5^2} \right)^2 \sigma_5^2, \quad (12.4)$$

where f_1 and f_5 are the L1 and L5 frequencies (1575.42 and 1176.45 MHz), respectively. If σ_1 and σ_5 are comparable then the ionosphere-free combination has roughly three times as much noise as either single frequency term, but is still substantially smaller than σ_{UIRE} . Furthermore, satellites do not need a grid correction to be used, thus satellites farther from the network and the IGP mask can be incorporated into the position solution. The dual frequency confidence bound for a single satellite is then given by

$$\sigma_{\text{tot_if},i}^2 + \sigma_{\text{fit},i}^2 + \sigma_{\text{iono_free},i}^2 + \sigma_{\text{trop},i}^2, \quad (12.5)$$

where σ_{air} is used in place of σ_1 and σ_5 in (12.4). The VPL term otherwise takes the same form as is used in today's L1-only system. However, now (12.5) is used in place of (12.1) to compute the uncertainty for each line of sight.

Several of the SBAS providers are evaluating plans to offer an L1/L5 service that makes use of these modernized signals. This upgrade will also be accompanied by the inclusion of other constellations as described below.

12.7.2 Multiple Constellations

In addition to GPS L5 development, there are several independent navigation satellite systems being developed with comparable civil frequencies [12.51]. Galileo is being developed by the European Union and is envisioned as being compatible with GPS in which each satellite provides ranging using signals covering the L1 and L5 frequencies with similar modulations. Although the system is still being deployed, it is envisioned that Galileo satellites will provide a service that is fully interoperable with the GPS civil signals.

In parallel, China is developing the BeiDou system whose signals are also planned to be compatible with GPS. Initially BeiDou has a B1 signal near L1 at 1561.098 MHz and another open signal, B2, at 1207.14 MHz. Beidou plans to provide signals at the L1 and L5 frequencies sometime after 2020 [12.52]. However, there is some uncertainty about the timeframe and exact nature of these new signals. Unfortunately, this uncertainty makes it difficult to develop the standards needed to create certified avionics. Offering signals precisely at L1 and at L5 does allow for the most convenient integration of BeiDou. However, signals at nearby frequencies could also be accommodated provided they are in aviation frequency bands.

The Russian GLONASS system has been operational for many years. Its current openly available signals are broadcast using different frequencies rather than different codes to distinguish the satellites. These frequencies range from ≈ 1598 –1605 MHz (near L1) and another open signal approximately between 1243 and 1249 MHz. There are modernization plans to broadcast signals at L1 and L5 that are more in alignment with the other constellations. It is not yet known when these new signals would be available.

EGNOS plans to correct both GPS and Galileo on both their L1 and L5 signals. SDCM is planning to augment both GPS and GLONASS. BDSBAS intends to augment both GPS and BeiDou. As these constellations mature and their signal offerings become better known, other SBASs may also choose to augment them. The additional signals essentially ensure that the user always has good geometry. With just GPS, one or more satellite outages can lead to a loss of vertical guidance service. However, with two or more constellations, service is tolerant to many satellite outages. Inclusion of multiple constellations into SBAS ensures continuity of service even if the constellations choose to maintain fewer satellites overall in the future. They also allow for the possibility of supporting even more demanding operations.

References

- 12.1 International Standards and Recommended Practices (SARPS), Annex 10 – Aeronautical Telecommunications (ICAO, Radio Navigation Aids 2006)
- 12.2 D. Lawrence, D. Bunce, N.G. Mathur, C.E. Sigler: Wide Area Augmentation System (WAAS) – Program status, ION GNSS 2007, Fort Worth (ION, Virginia 2007) pp. 892–899
- 12.3 Archive List of WAAS and SPS PAN Reports (Federal Aviation Administration, 2001–2016) <http://www.nstb.tc.faa.gov/DisplayArchive.htm>
- 12.4 E. Gakstatter: Using high-performance L1 GPS receivers w/WAAS for mapping/surveying, Proc. Stanf. Cent. PNT Symp. (2011)
- 12.5 K. Ali, M. Pini, F. Dovis: Measured performance of the application of EGNOS in the road traffic sector, GPS Solutions **16**(2), 135–145 (2012)
- 12.6 H. Cabler, B. DeCleene: LPV: New, improved WAAS instrument approach, ION GPS 2002, Portland (ION, Virginia 2002) pp. 1013–1021
- 12.7 R.G. Brown: A baseline RAIM scheme and a note on the equivalence of three RAIM methods, Navigation **39**(3), 301–316 (1992)
- 12.8 Minimum Operational Performance Standards for Global Positioning System/Wide Area Augmentation System Airborne Equipment (RTCA, Washington DC 2013)
- 12.9 J.A. Klobuchar: Ionospheric effects on GPS. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996), pp. 485–515, Chap. 12
- 12.10 L. Sparks, X. Pi, A.J. Mannucci, T. Walter, J. Blanch, A. Hansen, P. Enge, E. Altshuler, R. Fries: The WAAS ionospheric threat model, Proc. Beac. Satell. Symp., Boston (2001)
- 12.11 A. Komjathy, L. Sparks, A.J. Mannucci, A. Coster: The ionospheric impact of the October 2003 storm event on wide area augmentation system, GPS Solutions **9**(1), 41–50 (2005)
- 12.12 J.P. Collins, R.B. Langley: The residual tropospheric propagation delay: How bad can it get?, ION GPS 1998, Nashville (ION, Virginia 1998) pp. 729–738
- 12.13 K. Shallberg, P. Shloss, E. Altshuler, L. Tahmazyan: WAAS measurement processing, reducing the effects of multipath, ION GPS 2001, Salt Lake City (ION, Virginia 2001) pp. 2334–2340
- 12.14 R.E. Phelts, T. Walter, P. Enge: Characterizing nominal analog signal deformation on GNSS signals, ION GNSS 2009, Savannah (ION, Virginia 2009) pp. 1343–1350
- 12.15 C. Macabiau, C. Milner, A. Chabory, N. Suard, C. Rodriguez, M. Mabilieu, J. Vuillaume, S. Hegron: Nominal bias analysis for ARAIM user, ION ITM 2015, Dana Point (ION, Virginia 2015) pp. 713–732
- 12.16 K. Shallberg, J. Grabowski: Considerations for characterizing antenna induced range errors, ION GPS 2002, Portland (ION, Virginia 2002) pp. 809–815
- 12.17 S. Rajagopal, T. Walter, S. Datta-Barua, J. Blanch, T. Sakai: Correlation structure of the equatorial ionosphere, ION NTM 2004, San Diego (ION, Virginia 2004) pp. 542–550
- 12.18 T. Walter, A. Hansen, J. Blanch, P. Enge, T. Mannucci, X. Pi, L. Sparks, B. Iijima, B. El-Arini, R. Lejeune, M. Hagen, E. Altshuler, R. Fries, A. Chu: Robust detection of ionospheric irregularities, Navigation **48**(2), 89–100 (2001)
- 12.19 J. Blanch: Using Kriging to Bound Satellite Ranging Errors due to the Ionosphere, Ph.D. Thesis (Stanford Univ., Stanford 2003)
- 12.20 L. Sparks, J. Blanch, N. Pandya: Estimating ionospheric delay using kriging: 1. Methodology, Radio Sci. **46**(RS0D21), 1–13 (2011)
- 12.21 L. Sparks, J. Blanch, N. Pandya: Estimating ionospheric delay using kriging: 2. Impact on satellite-based augmentation system availability, Radio Sci. **46**(RS0D22), 1–12 (2011)
- 12.22 T. Walter, S. Rajagopal, S. Datta-Barua, J. Blanch: Protecting against unsampled ionospheric threats, Proc. Beac. Satell. Symp., Trieste (2004)
- 12.23 T. Walter: WAAS MOPS: Practical examples, ION NTM 1999, San Diego (ION, Virginia 1999) pp. 283–293
- 12.24 M. Grewal, P.-H. Hsu, T.W. Plummer: A new algorithm for WAAS GEO Uplink Subsystem (GUS) clock steering, ION GPS/GNSS 2003, Portland (ION, Virginia 2003) pp. 2712–2719
- 12.25 J. Vázquez, M.A. Sánchez, J. Cegarra, P.D. Tejera, P. Gómez Martínez: The EGNOS NOTAM proposals service: Towards full ICAO compliance, ION GNSS+ 2013, Nashville (ION, Virginia 2013) pp. 301–315
- 12.26 T. Walter, P. Enge, B. DeCleene: Integrity lessons from the WIPP, ION NTM 2003, Anaheim (ION, Virginia 2003) pp. 183–194
- 12.27 B. DeCleene: Defining pseudorange integrity – overbounding, ION GPS 2000, Salt Lake City (ION, Virginia 2000) pp. 1916–1924
- 12.28 J. Rife, S. Pullen, B. Pervan, P. Enge: Paired overbounding and application to GPS augmentation, Proc. PLANS 2004, Monterey (2004) pp. 439–446
- 12.29 T. Walter, J. Blanch, J. Rife: Treatment of biased error distributions in SBAS, J. Glob. Position. Syst. **3**(1/2), 265–272 (2004)
- 12.30 P. Enge: AAS messaging system: Data rate, capacity, and forward error correction, Navigation **44**(1), 63–76 (1997)
- 12.31 T. Walter, A. Hansen, P. Enge: Message type 28, ION NTM 2001, Long Beach (ION, Virginia 2001) pp. 522–532
- 12.32 L1 C/A PRN Code Assignments (US Air Force, Los Angeles Air Force Base 2016) <http://www.losangeles.af.mil/About-Us/Fact-Sheets/Article/734549/gps-prn-assignment>
- 12.33 FAA: <http://www.nstb.tc.faa.gov/>
- 12.34 A. Heßelbarth, L. Wanninger: SBAS orbit and satellite clock corrections for precise point positioning, GPS Solutions **17**(4), 465–473 (2013)
- 12.35 H. Manabe: Status of MSAS: MTSAT satellite-based augmentation system, ION GNSS 2008, Savannah (ION, Virginia 2008) pp. 1032–1059
- 12.36 T. Sakai, H. Tashiro: MSAS status, ION GNSS+ 2013, Nashville (ION, Virginia 2013) pp. 2343–2360

- 12.37 P. Feuillet: EGNOS program status, ION GNSS 2012, Nashville (ION, Virginia 2012) pp. 1017–1033
- 12.38 D. Thomas: EGNOS V2 program update, ION GNSS+ 2013, Nashville (ION, Virginia 2013) pp. 2327–2342
- 12.39 Monthly Performance Reports (European Satellite Services Provider, 2011–2016) https://egnos-user-support.essp-sas.eu/new_egnos_ops/content/monthly-performance-reports
- 12.40 J. Ventura-Traveset, D. Flament (Eds.): *EGNOS – The European Geostationary Navigation Overlay System – A Cornerstone of Galileo*, ESA SP-1303 (ESA, Noordwijk 2006)
- 12.41 R. Chen, F. Toran-Marti, J. Ventura-Traveset: Access to the EGNOS signal in space over mobile-IP, GPS Solutions **7**(1), 16–22 (2003)
- 12.42 GAGAN (Airports Authority of India, 2013) http://www.aai.aero/public_notices/aaisite_test/faq_gagan.jsp
- 12.43 India: GAGAN Implementation and Certification in India, 49th Conf. Dir. Gen. Civ. Aviat. Asia Pac. Reg. (ICAO, New Delhi 2012), http://www.icao.int/APAC/Meetings/2012_DGCA/
- 12.44 S. Karutin: SDCM Program Status, ION GNSS 2012, Nashville, TN 17–21 Sep 2012 (ION, Virginia 2012) 1034–1044
- 12.45 W. Song, J. Shen: China – Development of BeiDou Navigation Satellite System (BDS) – A program update, Proc. ION Pacific PNT, Honolulu (ION, Virginia 2015) pp. 208–236
- 12.46 Y. Yun: Influence of reference station distribution on the Korean SBAS performance, Proc. ION Pacific PNT, Honolulu (ION, Virginia 2015) pp. 964–969
- 12.47 S.S. Jan, W. Chan, T. Walter: ATLAB algorithm availability simulation tool, GPS Solutions **13**(4), 327–332 (2009)
- 12.48 T. Walter, J. Blanch, R.E. Phelts, P. Enge: Volving WAAS to serve L1/L5 users, Navigation **59**(4), 317–327 (2012)
- 12.49 T. Walter, P. Enge: Modernizing WAAS, ION GPS 2004, Long Beach (ION, Virginia 2004) pp. 1683–1690
- 12.50 R.S. Conker, M.B. El Arini, C.J. Hegarty, T. Hsiao: Modeling the effects of ionospheric scintillation on GPS/satellite based augmentation system availability, Radio Sci. **38**(1), 1–23 (2003)
- 12.51 C.J. Hegarty, E. Chatre: Evolution of the global navigation satellite system (GNSS), Proc. IEEE **96**(12), 1902–1917 (2008)
- 12.52 Z. Yao: BeiDou next generation signal design and expected performance, Int. Tech. Symp. Navig. Timing (ENAC, Toulouse 2015)

GNSS Part C Rec

Part C GNSS Receivers and Antennas

13 Receiver Architecture

Bernd Eissfeller, Neubiberg, Germany
Jong-Hoon Won, Incheon, Korea

14 Signal Processing

Jong-Hoon Won, Incheon, Korea
Thomas Pany, Neubiberg, Germany

15 Multipath

Michael S. Braasch, Athens, USA

16 Interference

Todd Humphreys, Austin, USA

17 Antennas

Moazam Maqsood, Islamabad, Pakistan
Steven Gao, Canterbury, Kent, UK
Oliver Montenbruck, Wessling, Germany

18 Simulators and Test Equipment

Mark G. Petovello, Calgary, Canada
James T. Curran, Noordwijk,
The Netherlands

Receiver Arch

13. Receiver Architecture

Bernd Eissfeller, Jong-Hoon Won

This chapter discusses the basic architecture of global navigation satellite system (GNSS) receivers. It starts with a breakdown of the receiver function into individual building blocks along the processing chain (front-end, down conversion, mixers, numerically controlled oscillators, correlators, tracking loops, data demodulation, navigation, user interface), and describes the respective functions. A dedicated section describes selected hardware solutions (example chipsets for front-end and baseband processing, offering different levels of integration and capabilities). Finally, receiver designs performing the signal processing in pure software as well as receivers based on configurable hardware are discussed.

13.1	Background and History	366	13.2.2	RF Front End	376
13.1.1	Analog Versus Digital Receivers	366	13.2.3	Analog-to-Digital Conversion	380
13.1.2	Early Military Developments	367	13.2.4	Oscillators	383
13.1.3	Early Civil Developments	368	13.2.5	Chip Technologies	386
13.1.4	Early Receiver Developments for Other Satellite Navigation Systems .	369	13.2.6	Implementation Issues	390
13.1.5	Early BeiDou Receiver Developments ...	371	13.3	Multifrequency and Multisystem Receivers	391
13.2	Receiver Building Blocks	372	13.3.1	Civil Receivers for GPS Modernization ..	391
13.2.1	Antenna	373	13.3.2	Galileo Receivers	394
			13.3.3	GLONASS Receivers	394
			13.3.4	BeiDou/Compass Receivers	395
			13.3.5	Military GPS Receivers	395
			13.4	Technology Trends	396
			13.4.1	Civil Low-End Trends	396
			13.4.2	Civil High-End Trends	396
			13.4.3	Trends in Military and/or Governmental Receivers	397
			13.5	Receiver Types	397
			13.5.1	Navigation Receivers Handheld	397
			13.5.2	Navigation Receivers Non-Handheld ...	397
			13.5.3	Engines, OEM Modules, Chips, and Dies	398
			13.5.4	Time Transfer Receivers	398
			13.5.5	Geodetic Receivers	398
			13.5.6	Space Receivers	398
			13.5.7	Attitude Determination Receivers	398
			References		399

In this chapter, the facts on global navigation satellite system (GNSS) receiver technology are presented. It provides a systems, engineering overview addressing all the different architectural elements. With respect to the main architectural subsystems, the possible technology areas are described and the implementation alternatives are traded-off with each other. Although the global positioning system (GPS) receiver is in the center of discussion, a wider view is given in order to cover also future satellite navigation systems (Galileo, GLONASS, BeiDou) and new signals and frequencies. Starting off

with historical reviews of civil and military receivers, early developments of receiver building-blocks are discussed with a sufficient level of detail: radio frequency (RF) front-end and antenna (including down conversion and filtering), analog-to-digital converter (ADC), crystal and other oscillators, RF and digital chip technologies, implementation issues, and multifrequency and multisystem receivers (including GLONASS, Galileo, military receivers). One goal of the chapter is also to identify the trends in receiver technology. Finally, the major types of receivers are basically described.

13.1 Background and History

GNSS receivers are now under development and production for over four decades. Already in 2013, around 1 billion civil GPS receivers (including 500–600 million GPS phones) were estimated to be in use on a global basis. The number of military or governmental receivers is about 300 000, in the North Atlantic Treaty Organization (NATO) and US-associated countries, which is much smaller. The long-term development cycle of receivers started off from 1975–1990 with analog boxes, and around 1990–1995, we saw the first integrated PC-type GPS original equipment manufacturer (OEM) cards. The next step was higher integration of these cards to *credit card* format till 2005. In contrast, mainly for the low-end market chipsets (RF chip + base-band chip), a single-chip hardware-based receiver was developed. In 2000, the software-defined radio (SDR) concept was applied to the GNSS autocorrelation receiver.

In general, the development line of GNSS receiver technology is strongly correlated with the history of the specific satellite navigation program and with progress in semiconductor technology. In the case of Navigation Satellite Time and Ranging (NAVSTAR) GPS, it started off in 1974 within the early hardware contracts for GPS Phase I [13.1]. In Phase I, three contracts for user equipment were established: Magnavox, Texas Instruments, and Rockwell International Collins Government Avionics Division for jamming resistant user equipment. These contracts included the development of monitoring receivers for the control segment, military receivers, and a civil type set as a utility for the military. Referencing again [13.1], the paradigm for the development of these receivers was: *Build a cheap set that navigates for less than 10000 US \$*.

In the early GPS user equipment development in the United States, we can identify mainly two classical do-

main: in the first domain, we see military GPS user equipment which was developed in different phases for land, aviation, and maritime applications, and in the second domain, a development of civil GPS receivers took place by stepwise higher integration of components.

13.1.1 Analog Versus Digital Receivers

Phase I receivers were mainly analog receivers (Fig. 13.1), which made use of a single-analog hardware channel, and later on a second hardware (H/W) channel was added. As in our days, these receivers had an L-band antenna and a low-noise amplifier (LNA). The channel as such was built up by a down converter to an intermediate frequency (IF) and made use of an analog correlator, which was implemented in the IF filter amplification stages (linear IF correlation). The digitization by the ADC was performed at the very end of the channel chain. The post-correlated signals, pseudorange, carrier-phase and/or Doppler measurements, and demodulated navigation data were processed in the digital domain with a navigation processor. For the tracking process, a sequential technique was used; a satellite was acquired, tracked for some time interval, and, after this was over, a switchover to the next satellite was done. Thus, parallel all-in-view tracking was not possible within this architecture. For a fast moving platform, a dead-reckoning device was necessary to aid the sequential procedure and to guarantee a navigation solution. Over the years, the slow sequencing (30 s per satellite) was replaced by faster multiplexing (switching with 5 ms per satellite) throughout the visible constellation.

In the 1980s, a technology change in the receiver architecture took place: it was quite obvious that the

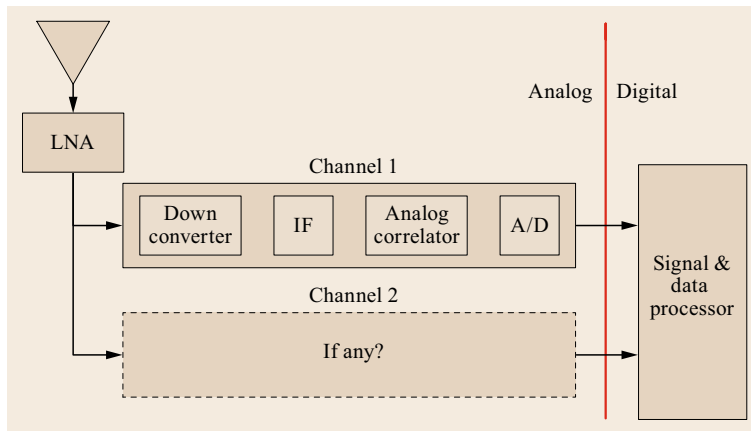


Fig. 13.1 Analog receiver architecture with sequencing and/or multiplexing

better solution for the GPS positioning problem should, on the one hand, be based on a parallel multichannel structure. This architecture made it possible to track five satellites at the same time or to track, for example, four satellites and use the fifth channel for acquisition or re-acquisition as well as for searching next a satellite in view. On the other hand, pioneering developments took place from 1977 to 1981 for an all-digital GPS receiver [13.2]. Receivers that made use of digital correlation via an application specific integrated circuit (ASIC) were in the market around the 1990s. At this time, it was usually possible to implement six channels on an ASIC-type correlator chip. Typical devices of this kind were the Magnavox MX 4200 [13.3] and the Novatel GPS cards [13.4].

In the data flow of a modern digital parallel receiver, the antenna, LNA, and down converter are still based on analog technology. After the down-conversion step, the digitization via the ADC takes place. After the ADC, the further signal handling gets digital, that is, a bank of digital channels is built up (Fig. 13.2). The classical design of a digital all-in-view receiver is to implement this bank of channels on an ASIC by use of a semiconductor process (e.g., complementary metal oxide semiconductor (CMOS), 90 nm). The ASIC is interfaced to a microprocessor (MP) which has the function to steer and control the ASIC and to read out the measurements and data.

13.1.2 Early Military Developments

In the area of early military GPS receivers, we mainly observe two development lines. One development went in the direction of a personal navigation device for ground troops. It started in 1980 with the Rockwell Collins Man-Pack (AN/PSN-8) which was a single-channel, dual-frequency, coarse/acquisition (C/A)-and precise code (P-code) receiver in the form factor of a backpack. The Man-Pack had a mass of 7.8 kg. One

decade later, the five-channel design was in production at Rockwell Collins and very large scale integration (VLSI) made significant progress. In the year 1993, the precision light weight GPS receiver (PLGR) was introduced in the NATO forces. The PLGR is a ruggedized handheld five-channel single-frequency L1, C/A-, and P(Y)-code receiver. The mass of the receiver is 1.2 kg, it has a jamming resistance of 24 dB, and anti-spoofing (A/S) is provided via the precise positioning service-security module (PPS-SM). The produced number of units is about 225 000. An upgrade with respect to an improved user interface and an improved power supply was provided with the PLGR+96. Since 2004, the PLGR was replaced by the DAGR (defense advanced governmental receiver) which is based on a GPS receiver application module-selective availability A/S module (GRAM-SAASM) form factor.

In the second development line, military aviation GPS receivers were designed. An early demonstrator of a five-channel military aviation receiver was the generalized development model (GDM) built by Rockwell Collins in contract with the Air Force Avionics Laboratory. The bulky system had a mass of 125 kg and was integrated on a flight-pad including the cooling system and the seats of the operators [13.5]. This activity was done in the concept demonstration Phase I of GPS. The first operational military aviation receiver was the Rockwell Collins 3A receiver [13.6]. This receiver was introduced in 1985 and was flown more or less on all Air Force, Navy, and US Coast Guard aircraft. Again it is a five-channel dual-frequency P-code receiver providing pseudoranges and delta ranges (integrated Doppler shift over 1 s) with a mass of 16.2 kg, a linear dimension of 42 cm and a power consumption of 116 W. The 3A was able to interface on the RF level with a controlled radiation pattern antenna (CRPA) and a fixed radiation pattern antenna (FRPA). Interestingly, besides other specific derivatives, a variant for Navy ship applications was derived which got the name 3S. Around 1990,

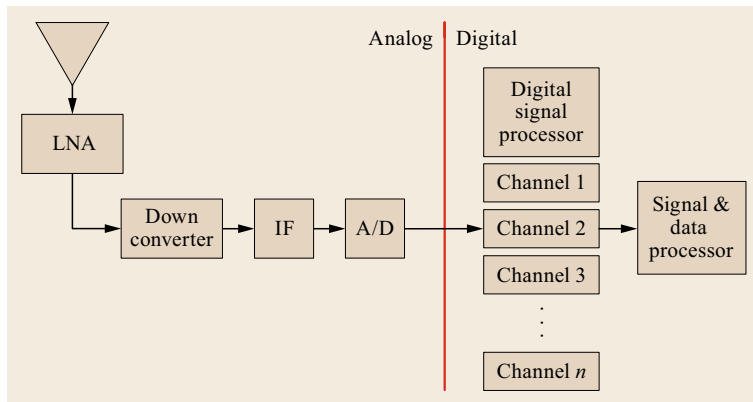


Fig. 13.2 Digital all-in-view receiver architecture

Rockwell Collins released the miniaturized airborne GPS receiver (MAGR) as a five-channel dual-frequency P(Y)-code receiver making use of a PPS-SM. Its software design is identical with the 3A receiver [13.7]. With 5.6 kg the MAGR has a smaller mass and a reduced form factor of $30 \times 17 \text{ cm}^2$ in the ground plane. Starting in 1998, a new contract for the production of the new MAGR 2000 was given to Raytheon in El Segundo. The MAGR 2000 is a form fit function compatible with the legacy MAGR. Since 2004, the MAGR 2000 is based on a standard electronic module, format E (SEM-E) form-factor GRAM-SAASM architecture.

13.1.3 Early Civil Developments

The first commercial GPS receiver was the Texas Instruments TI-4100. This receiver was a third-generation generation receiver and manufactured around 1981. It was based on large-scale integrated (LSI) components using the highest-speed bipolar digital technology which was available at this time. One version of the receiver was designed for commercial purposes, while the other was for a tri-agency version called GEOSTAR for use by the Defense Mapping Agency (DMA), National Oceanic and Atmospheric Administration (NOAA), and the US Geological Surveys (USGS).

The TI-4100 receiver (Fig. 13.3, [13.8]) is portable for field operations (it is about the size of a small suitcase) but can be mounted in a standard 19in rack if desired. It has a small lightweight handheld control display unit, and contains a keyboard and a display window. The receiver is modular in design and allows for easy removal and replacement of circuit boards. The size of the TI-4100 is $37 \times 45 \times 21 \text{ cm}^3$, the mass is 24 kg, and the power consumption is 93 W. The architecture of the receiver is based on a dual-frequency single hardware channel with a multiplexing tracking software package. It was able to track four satellites with a code and a carrier, that is, L1-C/A code, L1-P code, L1 carrier, L2-P code, and L2 carrier. If less than four satellites were visible, satellite data was acquired in a so-called degraded mode. In degraded mode, the receiver has the capability of adding an external atomic clock time standard input to provide clock-coasting.



Fig. 13.3 Texas Instruments TI 4100 receiver (University of the Bundeswehr, 1985)

For the civil scientific community in the 1980s, the TI-4100 was the only possibility to track the complete GPS Block I satellite signals with pseudorange and carrier phase on all the three codes. Thus, it played a very important role in pioneering of GPS methods and solutions, for example, in developing early techniques in precise carrier-phase positioning and geodetic applications.

On the Magnavox side in the GPS Phase I, the GPS PAC receiver and later on the GPS Phase IIB receiver were developed. Obviously, based on these first military developments in the early 1990s, the Magnavox MX 4200 digital receiver was commercially available. It was a six-channel L1 C/A-code highly integrated receiver, which gave access to the entire raw data structure on the C/A-code. This receiver was used by industry and many research institutes for bread-boarding activities and the demonstration of early applications. Regarding the NovAtel OEM cards, a basic innovation came up [13.4, 9], which was the introduction of a wide-band ($< 20 \text{ MHz}$) C/A-code receiver. This leads to the possibility of implementing a *narrow-correlator* technique in the signal processing. Previously, all C/A-code receivers [13.4] were mainly narrow-band (2 MHz) receivers, which employed a full single chip spacing in delay lock loop (DLL). The advantage of the narrow correlator was P-code-like thermal noise and better multipath rejection [13.9]. Besides these innovations, a new information policy on the receiver technology used in the GPS cards was established by NovAtel. This helped the civil user community to get a better insight into receiver design and signal processing and gave a push to university research in the field of receiver technology. The NovAtel GPS card of the mid-1990s was a 12-channel L1 C/A-code receiver. The 12 channels were implemented on two correlator ASICs (6 channels on each chip). Additionally, a transputer-based signal processor and an MP were used on the card.

At the beginning of 1990s, GEC-Plessey Semiconductors, a British-based International Electronics, Defense and Telecommunication Company, provided in the market a GPS L1 C/A-code receiver chipset solution, so-called GP2000 series, which was composed of three chips like GP2010/2015 for RF front-end, GP2021 for 12-channel digital correlator, and ARM 6/7 for the processor, together with all software source codes including detailed hardware design notes with a reasonable price [13.10]. Later, its digital parts (a digital correlator and a processor) were merged into an ASIC chip (GP4020) so that it became a two-chip solution. This GP series made it possible for many universities in the world to develop their own GPS L1 CA-code receiver solutions and/or to implement RF front-end devices being able to connect to general pur-

pose software-defined GPS receivers. The GPS chipset solution business part in GEC-Plessey Semiconductors was sold to other smaller companies in a series of merger and acquisitions, and became a technical baseline in many receiver developments.

In the next step of receiver development, the PC-type OEM board was miniaturized (higher integrated) on *credit-card* format from companies like, for example, Motorola, which had access to leading-edge semiconductor technology. A good example for this kind of receiver is the Motorola VP Oncore. There are many other examples for higher integration which happened in the early 2000s. This reduced form factor allowed the integration of small commercial handheld receivers and of small car navigation systems.

In general, it became obvious in the 1990s that low-end civil GPS receiver development at the low end would focus on L1 C/A-code only receivers, whereas for high-end users, availability of pseudorange and phase measurements on the second frequency was required. A dual frequency structure was necessary for compensation of the ionospheric path delay in differential and nondifferential GPS systems. Thus, a dramatic issue for the high-end community evolved, when it became clear that the US Department of Defence (DoD) intended to encrypt the public-domain precise code (P-code) to a classified Y-code. This was put into practice after the Gulf-War in 1991. Since 1994, all new Block II satellite made use of A/S encryption. Thus, no clear P-code would be available in the future, especially after the GPS full-operational capability (FOC) in April, 1995. Based on this perspective, all receiver companies, which developed for the high-end market such as Trimble Navigation, Ashtech, Leica, and NovAtel worked by reverse engineering on the problem to overcome the encryption of P(Y)-code L2. The techniques, still partly in use today, developed in this context were called codeless and/or semicodeless architectures. These techniques build on the fact that the P-code is encrypted by modulo-two addition of a W-code, which exhibits a notably lower chipping rate [13.11]. Thus, the basic idea [13.12] of advanced techniques is to narrow the RF bandwidth of 20 MHz down to the necessary smaller bandwidth [13.11] (500 kHz) by precorrelation of the received signal with a clear P-code. These methods are a generalization of the pure squaring, which goes back to the patent of *Charles Counselman III* [13.13]; this formed the basis for the Macrometer Model V-1000. With pure squaring the spreading code and the binary phase shift keying (BPSK)-modulated data is completely removed. A carrier-phase measurement resolving the half wavelength can be derived. Another approach made use by the fact that the Y-code is the same on L1 and L2. Thus, cross-correlating the signals

on two frequencies with each other allowed determining the path delay between L1 and L2. The Ashtech concept [13.11] got later well known under the term Z-tracking. A comprehensive overview and analysis of many semicodeless methods is given in [13.14].

13.1.4 Early Receiver Developments for Other Satellite Navigation Systems

It can be seen from the history of GPS receivers that the step into GNSS receiver development is initiated if the satellite navigation system under development has obtained a certain state of maturity. Usually, the first types of receivers which are implemented are receivers for monitoring or sensor stations. Also, some early test user receivers are developed by the investment of research and development money. The investment in the GNSS receiver development by the existing or newly to-be-formed industry always happens with a delay of years. The reason is that, on the one hand, the stability of a satellite navigation program has to be shown by the service provider; on the other hand the compatibility with the claimed performance requirements has to be demonstrated.

Early GLONASS Receiver Developments

It is known that the Russian Global Navigation Satellite System (GLONASS) was developed in 1972 in parallel to GPS, but little is known about early GLONASS receiver development. The first GLONASS receivers like the ASN-16 or the Skipper became available in Europe around 1990 when Russian space industry opened to the Western world in order to establish joint-venture contracts with international space industry. German companies, that is, Kayser-Threde, Aerodata, and MAN Technology had several meetings with Russian colleagues. Mainly two Russian institutions were involved: the Institute of Space Device Engineering (ISDE) and Moscow and the Leningrad Scientific Research Radio Technical Institute (LSRRI), St. Petersburg. Contacts between the USA and Russia in this context are also reported. As the result of the German-Russian consultations, two types of receivers were available in Europe since 1992. The first one was the aviation receiver ASN-16 and the second one was a marine receiver called Skipper.

The ASN-16 (Fig. 13.4) was a single-channel GLONASS receiver which consisted of several subunits: a control and display unit, a navigation computer, an RF unit, an antenna, and a pre-amplifier. Via an ARINC-429 interface, raw data could be sent to a PC for display, storage, and further processing. The specified weight for the receiver was 25 kg and the power consumption was on a level of 180 W. In 1992, it was

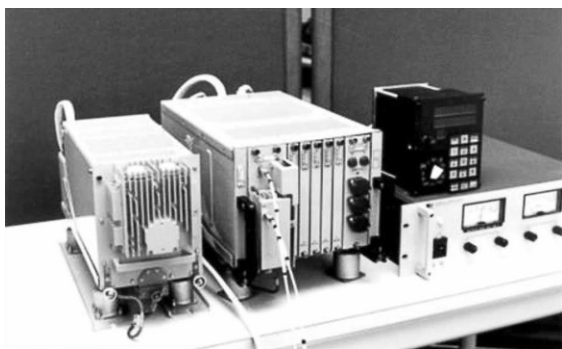


Fig. 13.4 Russian GLONASS Receiver ASN-16 (courtesy of OHB)

announced by the ISDE to proceed with a higher integrated aviation receiver which was called the GNOM that was a 5 kg and 30 cm unit. In this weight class, the ISDE demonstrated a geodetic type of receiver named REPER, which was designed to provide precise carrier-phase measurements with 1–3 cm accuracy. Different GLONASS tests with the ASN-16 are reported in [13.15].

Early Galileo Receiver Developments

In March 2002, the Galileo development phase was launched by a decision of the EC council. The official kick-off of the Galileo Phase C/D which marks the start of the industrial development goes back to December, 2004. In the time frame of the early 2000s, the first activities concerning the development of Galileo or combined Galileo/GPS receivers took place. These receiver development projects were funded mainly by European Space Agency (ESA). Later on, additional funding by European Commission, for example, in the framework program, FP6 (2003) and the FP7 (2007) context was provided for receiver activities.

For the Galileo ground mission segment (GMS) as well as verification and test activities of the payload, multifrequency receivers have been developed in Europe within ESA contracts.

The Galileo ground reference receiver (GRR) has been under development since 2002 [13.16]. It forms a part of the ground reception chain (GRC) which is a basic element of the Galileo GMS. Initially, Thales Alenia Space Italia (TAS-I), NovAtel, and Space Engineering were selected by the European Space Agency (ESA), its prime contractor European Satellite Navigation Industries (ESNIS), and Thales Alenia Space France (TAS-F), and work on the program began in June 2005. TAS-I has awarded NovAtel milestone contract for the continued development of the GRC reference receiver [13.17]. In the joint venture, 25 production units were to be delivered in order to support

the in-orbit validation (IOV) phase of Galileo. NovAtel is delivering receiver components, which are integrated into the GRC.

The Munich-based company, IfEN, awarded a sub-contract from Siemens Austria in the Galileo program to develop and produce a set of payload test receivers. The receiver is used to characterize the transmitted signal of a payload generator or constellation simulator before launch. The receiver is usually directly connected to the payload by use of an attenuation component. Mainly two versions of the payload test receiver were built. The first-generation IfEN Rx payload test receiver for IOV was built. It supports all services on Galileo IOV, for example, public regulated service (PRS) via an interface to a code encryption unit. The project was scheduled from 2006 to 2012. Additionally, a payload test receiver was developed and built in the time frame of 2010–2013 for the Galileo FOC testing. This receiver is a third-generation Rx of IfEN. It supports all services on Galileo FOC 2010–2013 including the PRS. It should be mentioned that the IfEN receiver development activities were initiated by the realization of the Galileo test range GATE, which was funded by German Space Management (DLR) (Bonn). For GATE, the first-generation E1, E6, E5a, E5b receiver was developed. This receiver was upgraded to a second-generation receiver called the NavX-NTR (Fig. 13.5).

In the field of test receivers for Galileo development phases, a Galileo experimental test receiver (GETR) was built by the Belgium-based company, Septentrio, in 2005: the GETR was used for the in-orbit validation of Galileo signals transmitted by GSTB-V2 satellites [13.19].

Other important developments of Galileo receivers or combined GNSS receivers in Europe are the test user receivers (TURs). A basic TUR was developed by Septentrio till 2010. The TUR is an all-in-view multifrequency Galileo/GPS receiver capable of tracking all Galileo signals together with GPS L1 and L5. The idea of the TUR project is to provide a user-like test receiver for the assessment of, for example, the user equivalent range error (UERE) of Galileo. A PRS-capable version (TUR-P) was developed in a joint venture between



Fig. 13.5 Multifrequency Galileo test receiver NavX-NTR of IfEN GmbH (after [13.18], reprinted with permission)

Septentrio and the UK-based company, QuinetiQ, in 2010. In contrast to the Septentrio TUR development, a TUR-N and a TUR-P are under development at the Thales group.

A prototype Galileo PRS receiver called the Bavarian security receiver (BaSE) is developed under the lead of the Fraunhofer Institute for Integrated Circuits [13.20]. This is the first initiative in Germany to utilize the PRS of Galileo.

Besides these agencies-funded Galileo receiver projects (since the Galileo signal interface control document (ICD) was released in 2006), the open Galileo signals were integrated by many manufactures on a worldwide basis in the low-end and the high-end receiver domains. Thus, many chips are now on the market which provide a Galileo capability besides GPS.

13.1.5 Early BeiDou Receiver Developments

The GPS development program of US DoD and their announcement in 1984 to open it to civilian aircraft for free of charge served as the stimulus for China to begin to explore research on satellite-based positioning technology in 1985; the national-level satellite-based navigation system development program in China was named BeiDou (a Chinese name for *Big Dipper*). Later, this development program was extended to the regional satellite navigation system (BeiDou-1) to cover China and neighboring regions composed of two separate constellations operating since 2000. And, a full-scale global satellite navigation system (also known as COMPASS but recently named as BeiDou-2 due to frequency compatibility issue against Galileo) is currently under construction to have 35 satellites; it became operational in the Asia-Pacific region since 2011 with 10 active satellites, and its target year is 2020+.

Regarding BeiDou receivers, in 1995, the National University of Defense Technology in China began to investigate on baseband signal-processing algorithms. In 1998, the equipment is being tested in the ground station in Beijing. According to the data on the screen, the system was running well within the theoretical limits. It took three years in total from the theoretical design to the production of the receiver.

In November 2001, BeiDou became the third satellite-based navigation and positioning system together with existing GPS and GLONASS. However, problems occurred when the system was put into practice; the size of the receiver was too large (backpack-size). In 2004, the first BeiDou-1 portable navigation device (PND), which solved the portability problem of the receivers, in China was released. This means that it could be applied in many fields such as safety of life, defense, aviation, and so on. And it could also be

used to communicate between different receivers to get the positions of each other by the short message functionality which is unique for the BeiDou system. In May 2008, a great earthquake took place in WenChuan, SiChuan Province, China. After 5 days of this disaster, the Beijing Satellite Navigation Center (BJSNC) received a message from the place where the disaster happened, indicating that the after-shock would keep happening in BeiChuan, and the water level of Haizi would keep raising. This message was delivered by the BeiDou-1 receiver owned by the rescue soldiers. This is the most famous application of the BeiDou-1 receiver in the public. In 2007, when one of the BeiDou satellites was running, a serious jamming was detected, causing the interruption in receiving the signals. After the analysis of the data by experts, it was discovered that the complex electromagnetic environment caused this interruption of the signal. And the research on antijamming began. In May 2008, antijamming equipment was successfully developed.

Currently in China, there are tens of companies researching the BeiDou chip-manufacturing process. However, only a dozen of companies have successfully developed the BeiDou chips; several of them own a complete chipset solution: the RF chip and baseband chip. For example, in the fleet management system for the government-owned vehicles of Guangzhou city, baseband chips from the Unicore Communications, Inc. and RF chips from the Guangzhou Ruixin Information Technology Co. are used. And, on the buses and trucks operated by the Ministry of Communications, more than 80 000 receivers were installed. Other famous company in this field includes the OLinkStar Co., the HuaXun Microelectronics Inc., the Beijing Hwa Create Corporation Ltd, the Hangzhou Zhongke Microelectronics Co., the Shanghai Fudan Microelectronics Group Co., and so on. The BeiDou baseband and RF chips in the market are independently designed. Chips from the worldwide major brands are integrated together; in this technique, it is claimed that there is a big gap (at least 15 years) between the domestic companies in China and major worldwide brands. Also, there is still a large gap in the manufacturing technique; the domestic companies can reach the 55 nm level while the major brands can reach a level below 40 nm. And until now, there has been no successful story of a Chinese companies to provide a single-chip combo solution, integrating the navigation chip and the communication chip in the largest market of navigation applications, for example, the mobile market. In the field of high-accuracy boards, the representative companies in China are Hi-Target Survey Instruments Company Ltd, the ComNav Technology Ltd, the South Survey/Mapping Instrument Ltd, the CHC Co., the Unistrong Co., and so on.

The followings are examples of BeiDou chips, receiver modules, and various application receiver systems compatible with GPS, GLONASS, or Galileo. UB240 is a BeiDou/GPS dual-system quad-frequency OEM board developed by Unicore/BDStar based on its multisystem, multifrequency, high-performance system-on-a-chip (SoC) technology with a self-owned intellectual property. UB240 uses low power consumption design and offers millimeter-level carrier-phase observations and centimeter-level real time kinematic (RTK) positioning precision, and is in support of multipath mitigation. Advanced technology for a long distant RTK is particularly appropriate for the application of high precision measuring and positioning [13.21]. CC50-BG was developed by OLinkStar based on their ProGee II GNSS engine. It provides the GPS, GLONASS, and BeiDou combined navigation solution. OLinkStar announced that this is the smallest satellite navigation module with BeiDou positioning functionality covering all over the world. The CC50-BG has been widely used in vehicle navigation and handheld devices on the Chinese market [13.22]. National

University of Defense Technology recently developed a dual-band GNSS receiver with low-IF architecture for high-powered location-based service (LBS) in a 55 nm CMOS process. The receiver embodies two independent IF channels to support GPS-L1 and Compass-B1 signals, and can provide three operating modes (GPS-L1 and Compass-B1, GPS-L1, Compass-B1) with either passive or active antenna. The K508 GNSS receiver board is the first eight-frequency (BeiDou B1/B2/B3, GPS L1/L2/L5, GLONASS L1/L2) OEM board developed by the ComNav Technology Ltd that owns full independent intellectual property rights of this receiver board. It uses fast BeiDou high-dynamic and -accuracy computation algorithm engine to achieve a decimeter-level accuracy for the baseline of hundreds kilometers in a single epoch. The size, interface, and data command are compatible with OEM boards from major brands that support the same eight-frequency signals. It can be used in the areas of multisystem, multifrequency reference station, high accuracy surveying, aerospace applications, deformation detection, machine control, defense, etc. [13.23].

13.2 Receiver Building Blocks

A conventional GNSS hardware receiver is basically composed of the analog part, the digital part including the application processor, and the interfaces for input–output. The following building blocks [13.9] may be identified: single- or multifrequency L-band antenna plus cable, RF front-end including the LNA, the oscillator, down converter and mixers, and the bandpass filters, the ADC plus, optionally, an amplifier gain control (AGC). The digital part consists of a baseband integrated circuit containing the correlators, the MP, and the read only memory (ROM) and random access memory (RAM) memory units. Usually, correlator ASICs are in use plus a separate MP and memory elements. Sometimes these elements are integrated onto a single chip (systems-on-a-chip: SoC). Basically, the functions of the building blocks may be briefly described as follows:

- The antenna is to set the gain toward satellite (± 3 dBic) and is essential for out-of-band interference rejection.
- The preamplifier (LNA) is to set the noise figure F .
- The surface acoustic wave (SAW) filter is to band-limit the signal (bandwidth has impact on accuracy, for example, correlation loss).
- The ADC converts analog signal to digital (n-bit ADC).

- The AGC maintains the signal-plus-noise magnitude within the limits required by ADC level detectors.
- Major signal-processing functions are the following:
 - Reference signals (codes/carriers) are locally generated.
 - Autocorrelation is de-spreading the signal (high-frequency part removed).
 - Tracking loops (DLL and phase lock loop (PLL)) necessary to follow the time variations of the signals.
 - Carrier-aiding used to reduce noise in DLL, provides also dynamics.
- The reference oscillator generates fundamental frequency.
- The application processing computes the user's position, velocity, and time.

Before describing in more detail the different building blocks, some terms should be clarified which are used in the satellite navigation community to classify a type of receiver. In a classical hardware receiver, the baseband operations are hardwired on a baseband chip or an integrated circuit by making use of a state-of-the-art semiconductor integration process. The degree of flexibility in the case of upgrades is small. In an SDR,

hardware and software technologies are used in parallel to generate a reconfigurable solution. The goal is the ability to update the system by software changes. In any case, an analog front-end down to the ADC will be present: on the processor side, an field programmable gate array (FPGA), an MP or a general purpose computer or a mix of such a processor could be used. In GNSS software receivers, a general purpose processor is used: e.g., an Intel PC or an advanced risk machine (ARM) processor. All signal processing is programmed in a higher programming language like C/C++.

Figure 13.6 shows the building block diagram of a GPS C/A-code receiver. The specific typical technical parameters for L1 C/A-code receivers are added. In general all GNSS receivers are designed in such a way. For a more general GNSS receiver, the receiver parameters have to be adjusted to the specific signal structures.

13.2.1 Antenna

The role of the antenna (Chap. 17) is to receive an electromagnetic wave and convert it to an electronic signal. The modern goal in satellite navigation is to work with an all-in-view constellation. Thus, all satellites above the horizon should be received on the L-band carrier frequencies with an appropriate bandwidth. Therefore, a hemispherical radiation pattern is required for simultaneous L-band signal reception. In a real scenario, the maximum gain (+3 dBic) of the antenna is usually in the bore-sight direction (at 90° elevation). However, a significant gain should be present at lower elevation angles, because lower elevation signals suffer from higher attenuation (free space loss and atmospheric loss). The 0 dBic point is approximately located at 70°. However, below 5° of elevation, a sharp pattern roll-off is necessary in order to attenuate multipath signals which are reflected from the ground or the platform mounting surface. In the lower hemisphere of the an-

tenna pattern, high attenuation below -8 dBic or more is required (small back-lobes). Additionally, the antenna has to work with polarized signals: It should provide a good gain for right-handed circular polarized (RHCP) waves and a low gain for left-handed circular polarized (LHCP) waves. A sample gain pattern is illustrated in Fig. 13.7.

To understand polarization is essential for GNSS: in electromagnetics [13.25], the polarization state of a wave (and antenna) is described by its location (spherical coordinates) on the Poincare sphere (Fig. 13.8). Recall, that RHCP was introduced in GPS to get independent of the antenna pattern orientation at the satellite and the receiver. Circular polarization also plays a role in multipath rejection. In [13.25], a so-called wave-to-antenna coupling factor η_d is introduced

$$C_d = \eta_d C$$

$$\eta_d = \left(\frac{1-d}{2} \right) + d \cos^2 \left(\frac{\text{MM}_a}{2} \right)$$

$$d = \frac{\text{polarized power}}{\text{total power}} \quad (0 \leq d \leq 1) . \quad (13.1)$$

In these equations, C_d is the signal power at antenna output, C is the signal power at antenna input, η_d is

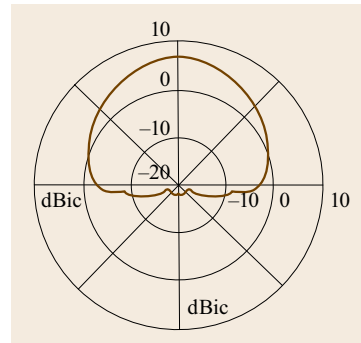


Fig. 13.7 Typical RHCP antenna gain pattern for a GPS patch antenna (after [13.24])

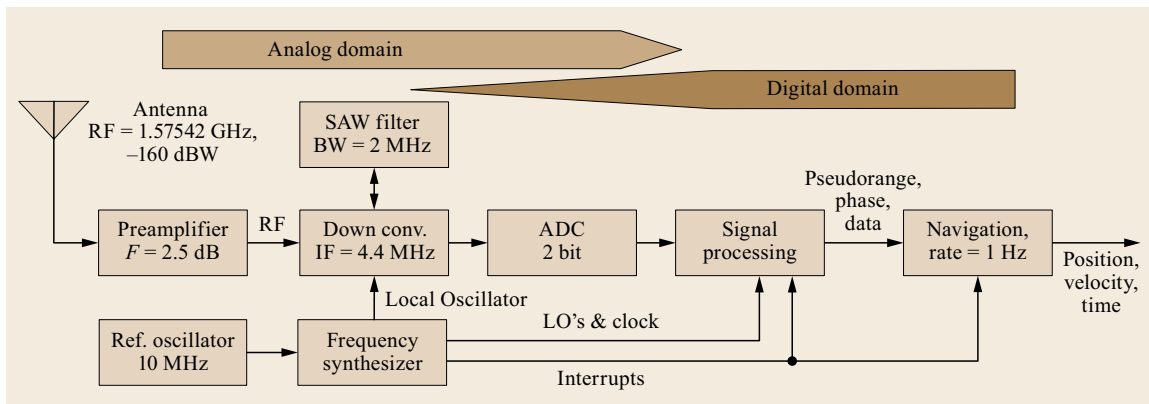


Fig. 13.6 An example of building blocks for a GPS C/A-code L1 receiver

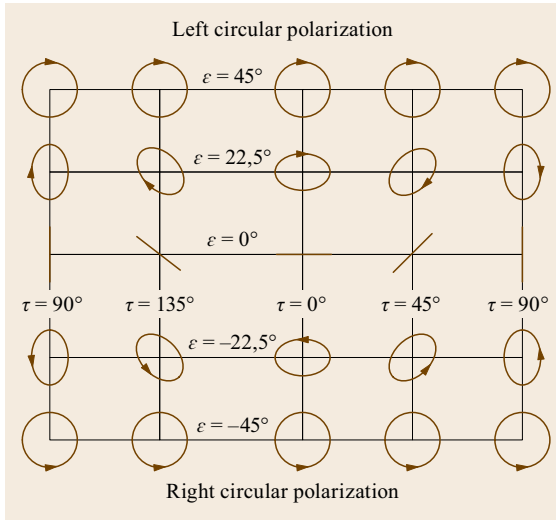


Fig. 13.8 Polarization states on the Poincare sphere (after [13.25]) (latitude = 2ε , longitude = 2τ)

the wave-antenna coupling factor, d is the degree of polarization ($d \approx 1$ for GPS), and MM_a is the spherical angle between the polarization states of the wave and of the antenna on the Poincare sphere. For example, the angle MM_a is 180° for an RHCP antenna and an LHCP wave, while it is 90° for an RHCP antenna, and a horizontal or vertical linear polarized wave. Right-handed and left-handed elliptical polarizations have intermediate locations on the Poincare sphere.

For a GNSS antenna five primary design parameters may be identified:

1. Center frequency
2. Bandwidth
3. Radiation pattern and gain
4. Axial ratio (AR)
5. Phase response.

An antenna behaves like a bandpass filter around the center frequency. Thus, the bandwidth should be large enough to pass the modern GNSS wide-band signals. The axial ratio ($AR \geq 1$) is another parameter to describe the polarization state. It determines the latitude (2ε) on the Poincare sphere by the relation $2\varepsilon = \cot^{-1}(\pm AR)$. The phase response is important for precise carrier-phase tracking because it is related to the stability of the phase center in spatial directions.

In general, four physical design forms of antennas are in use [13.26, 27]. It should be outlined that the antenna design chosen is related to the target application: civil, military, hand-held, mobile phone, aviation, space, and geodesy.

Active and Passive Antennas

A technical difference has to be made between active and passive GNSS antennas. In an active antenna, the LNA is integrated into the antenna. The advantage is that the cable loss can be overcome by applying an appropriate gain at the LNA output. The active antenna needs a remote power feed from the receiver via the antenna cable. In a passive GNSS antenna, the LNA is built-in into the RF unit of the receiver. Usually, this approach is only used if the distance between the antenna element and the receiver is short (some cm). The market is dominated by active antennas.

Helix Antennas

If properly designed, helical antennas can be the best antennas for GNSS reception. Helix antennas appear in GNSS in two forms: spiral helix and/or quadrafilary helix (volute). Helical antennas provide wider bandwidths than the planar antennas at the expense of higher back-lobes. The bandwidth of a broadband helix antenna can be sufficient for receiving wide-band signals up to 30–50 MHz. Phase-center stability is an issue. Using multiple-turn techniques [13.26], the bandwidth of the helical antenna can be increased. This can be helpful in order to cover the dual frequencies L1 and L2 of the GPS by a single antenna. The radiation pattern characteristics and the symmetry of the pattern symmetry are improved. The same holds for the axial ratio. The remaining problem of poor pattern roll-off in the horizon and high back-lobes has to be overcome by means of the ground plane, choke-ring, or the use of a stealth ground plane (with absorbing material). In early geodetic receivers like the TI-4100, a helical antenna was used. As described in [13.27], it is possible to build small ($18 \times 10 \text{ mm}^2$) helical antennas by dielectric loading.

Planar Antennas

The main technology used for low-end GNSS applications which require small antenna profiles [13.26] is the planar antenna technology. Planar antennas may be produced with medium to low cost. Planar technology is usually implemented as a classical microstrip patch antenna or a small microstrip version. We will also cover ceramic chip antennas in this context. Microstrip antennas [13.27] have a thickness of some mm and can be integrated on dielectric substrates. A patch antenna yields sufficient AR, gain pattern, and stable phase-center location and adequately low back-lobes. The a priori limiting factor of a microstrip antenna is the relatively narrow bandwidth of around 3–5 MHz. So-called wide-banding concepts have to be used in order to enhance their bandwidth.

For many low-end applications like mobile phones or personal navigation devices, the size of the GNSS

antenna is the critical factor. For a quadratic patch, the engineering rule for the size D is given by

$$D = \frac{\lambda}{2} \frac{1}{\sqrt{\epsilon_r}}, \quad (13.2)$$

where λ is the free-space wavelength in L-band, that is, about 200 mm and ϵ_r is the dielectric constant of the used substrate. This leads immediately to the fact that a microstrip antenna element with $\epsilon_r = 2.0$ has a size of $50 \times 50 \text{ mm}^2$. Using alternative substrate materials like Al_2O_3 , the dielectric constant could be increased to $\epsilon_r = 10$ or even higher for SrTiO_3 with $\epsilon_r = 270$. The latter material would theoretically allow designing a ceramic L-band chip antenna with a size of only $3 \times 3 \text{ mm}^2$. A small ground plane and circuitry have to be added.

Microstrip antenna elements are also the standard choice for building antenna arrays.

Array Antennas

Controlled radiation pattern antennas (CRPAs) are in use for decades in military applications. An analog version was the GPS Antenna System 1 (GAS-1) which basically consisted of a seven-element array on the associated ground plane. The physical space between a pair of antenna elements is $\lambda/2 = 10 \text{ cm}$. The design resulted in a large aviation antenna with a diameter of 35.6 cm, a height of 2 cm, and a mass of 1.8 kg.

Multiantenna arrays can be used on the one hand for the purpose of beam forming in the directions of GNSS satellites. Here, the purpose is to increase the gain and to enhance the carrier-to-noise ratio (C/N_0) of specific satellites. On the other hand, the more classical application is the spatial nulling of jammers in the horizon. With n array elements, $n - 1$ pattern nulls may be generated, that is, $n - 1$ jammer or interference sources can be eliminated. A simple four-antenna array is illustrated in Fig. 13.9.

Depending on the required performance, synthetic or adaptive techniques can be used to perform the beam forming and/or nulling. It can be carried out by means of analog techniques (in RF or IF frequency domain) or digital techniques.

The use of an antenna array compared to an omnidirectional antenna imposes additional challenges to be addressed, if the user is changing the attitude of the antenna array and the variations in the received signal have to be monitored, and, in the case of a phased array, the beams have to be controlled in their direction to the satellites. Depending on the size of the array, significant complexity is added to the receiver architecture and signal processing because a complete receiver chain is necessary behind each array element. For high-

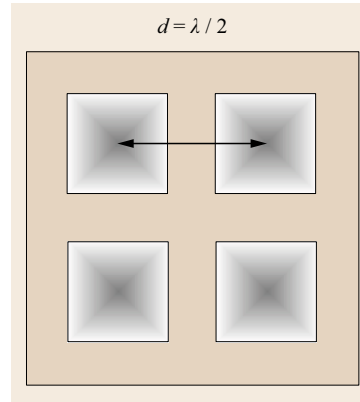


Fig. 13.9 Principle of array antenna with four planar elements

precision carrier-phase application, the determination of the antenna phase centre is critical because of the time-dependent variation of the antenna gain pattern.

CRPAs play a role in jamming resistant GNSS user equipment. Because a CRPA has a huge form factor and profile, it is only usable on larger platforms (large military–civil transport aircraft, ships, static reference stations on ground). Because of higher complexity, it adds also significant cost to the GNSS receiver. The future challenge will be in the design of small-sized arrays [13.28] by making use of ultra-small patches with dielectric loading. For the more complex RF circuitry, single-chip solutions with multiple RF chains are already available.

Additional Elements for Multipath Suppression

In high-precision carrier-phase applications, for example, for surveying and geodesy, additional requirements exist for the antenna: ground reflections and the propagation of surface waves have to be suppressed, whereas on the other hand, the antenna phase center has to be determined with submillimeter precision. Fortunately, the technical solution for the first problem area supports also a stable phase-center. Ground reflections enter into the antenna element as multiple reflections. These can be treated by the fact that a single reflection changes the polarization state of the reflected wave to LHCP and is thus mitigated by an RHCP GNSS antenna. In any case, a change of the polarization state of the GPS wave will happen only with perfect conductors and perfect specular reflectors.

In order to generate a sharp gain pattern roll-off at low elevation angles, three solutions exist which are commercially available by high-end receiver manufacturers.

The first solution that is known since the 1980s is the use of co-centric choke rings around the antenna element. The choke rings have a depth of $\lambda/4$. The concept is to generate an area of high impedance which then prevents propagation of surface waves [13.29].

A second solution introduced by Trimble [13.15] is to use a Stealth ground plane. A material is used which offers a high radially increasing sheet resistivity from the antenna element to the edge of the ground plane. Similar technology is used in the Stealth military aircraft systems. The performance of the antenna is said to be independent of the carrier frequency. Thus, the technology could be a good choice for multifrequency, multisystem, high-end antennas.

A third solution is the *pin-wheel* technology invented by NovAtel. In this technology, printed rings are used instead of massive physical choke rings.

Phase Response and Phase Center Variation

The phase response of a GNSS antenna varies with frequency, elevation angle (and azimuth), and temperature. In a good antenna design, the phase-response uncertainty has to be minimized. The idealized intersection point of the electromagnetic wave from different satellites is the antenna phase center. Depending on the antenna design, the phase center may exhibit offsets of up to 10 cm in height and mm to cm level horizontally relative to the mount point. The phase-center variation, which describes the deviation of the waveform from a sphere, adds additional direction-dependent biases of several mm to cm to the carrier-phase measurement and must be taken into account in the high-precision modeling of GNSS observations (Chap. 19). Calibration techniques for determining the phase-center offset and variation of receiver antennas are, furthermore, described in Sect. 17.6.2 of this Handbook.

13.2.2 RF Front End

In a modern GNSS receiver, the RF unit is the most critical subsystem, because it determines cost, size, and power consumption of the receiver. Especially for highly integrated single-chip receivers, the front-end design is of primary importance. The overall bandwidth of the pre-correlation filter chain determines the post-correlation pseudorange accuracy, because a trade

off decision between wide- and narrow-band design is done. The key components of the RF front end are the LNA, the RF, and IF filters, and also the local oscillator (LO) used for down conversion. This RF oscillator is usually coupled to a crystal oscillator by means of a phase-lock loop (PLL). The down-conversion scheme of a GNSS receiver poses requirements on the crystal oscillator (stability, phase noise, and frequency).

Concepts of RF Processing

The block diagram for a typical GNSS receiver front end is shown in Fig. 13.10. The first necessary subsystem in the chain for such a front end is the antenna, which was already described in the previous chapter. After the antenna, usually an RF filter is used to eliminate unwanted emissions that could otherwise enter the LNA. The necessary selectivity of this filter determines the technology and also (together with the LNA noise figure) the loss introduced into the pass band.

As stated in the previous chapter, most antennas are built as active antennas, that is, they include the LNA and the RF pre-filter. After the LNA, a second RF filter, which could provide the necessary image rejection, is used. The RF signal is converted down into a first IF in the range of typically hundred MHz. An IF filter, which could be much narrower than the RF filter, is used to eliminate unwanted spectral parts. A second down conversion delivers the received signal in a frequency range of several MHz. In a real receiver scenario, the center frequencies of the front-end depend on the implemented frequency plan. The subsequent IF filter eliminates unwanted spectral components. It ensures that the ADC fulfills the sampling theorem.

Mathematical Description of Front-End Operations

Descriptions of the front-end functions are found in several textbooks, for example, [13.9, 11]. In this section, we perform a description in the frequency domain based on Fourier transform (FT), making use of the Dirac delta function and the frequency-shift theorem of

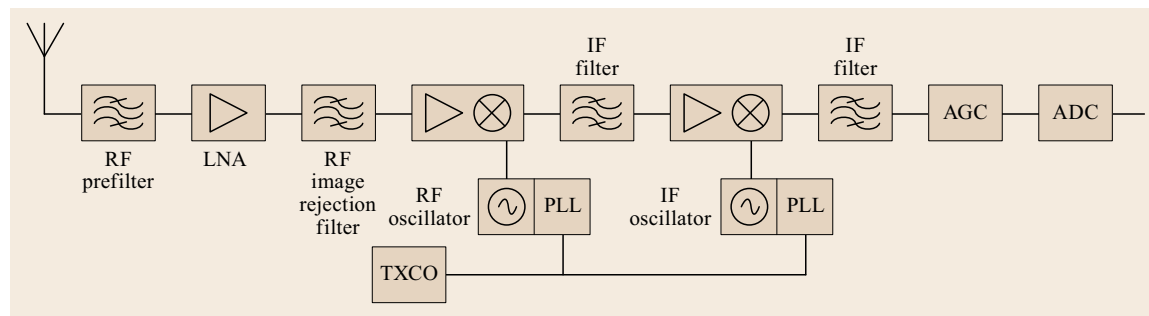


Fig. 13.10 Typical heterodyne RF front end of GNSS receivers (after [13.30])

Fourier transform. Because the carrier frequency is substantially larger than the chipping rate of spreading code (e.g., by a factor of 154 for L1 P-code), we consider the code chip $c(t)$ constant and discuss only an unmodulated, that is, constant amplitude carrier

$$s(t) = a \cos(\omega_1 t + \phi) + n(t), \quad (13.3)$$

where $s(t)$ is the signal, a is the amplitude, $\omega_1 = \omega_0 + \Delta\omega_D$ is the Doppler-shifted nominal carrier frequency with ω_0 denoting the nominal carrier frequency, $\Delta\omega_D$ is the Doppler shift, ϕ is the phase of the carrier, and $n(t)$ is the thermal noise. We define the Dirac delta function as usual

$$\delta(\alpha) = \begin{cases} \infty & \text{for } \alpha = 0, \\ 0 & \text{for } \alpha \neq 0. \end{cases} \quad (13.4)$$

The Fourier transform of the harmonic signal $s(t)$ may then be written as

$$\begin{aligned} F[s(t)] &= S(\omega) \\ &= a\pi[\delta(\omega + \omega_1) + \delta(\omega - \omega_1)] + N(\omega) \end{aligned} \quad (13.5)$$

with a spectral representation of signal $S(\omega)$ and noise $N(\omega)$. For the mixing operations, we multiply the signal (13.3) with the mixing signal

$$\cos(\omega_{LO})t = \frac{1}{2}(e^{+j\omega_{LO}t} + e^{-j\omega_{LO}t}) \quad (13.6)$$

in the time domain. The frequency ω_{LO} is the reference frequency of the LO generated by the PLL of the mixing stage. Applying the frequency-shift theorem,

$$F[g(t)]e^{j\omega_0 t} = G(\omega - \omega_0) \quad (13.7)$$

yields

$$\begin{aligned} S_{\text{mix}}(\omega) &= F[s(t) \times \frac{1}{2}(e^{+j\omega_{LO}t} + e^{-j\omega_{LO}t})] \\ &= \frac{1}{2}S(\omega - \omega_{LO}) + \frac{1}{2}S(\omega + \omega_{LO}) \end{aligned} \quad (13.8)$$

and by computation,

$$\begin{aligned} S_{\text{mix}}(\omega) &= \frac{a}{2}\pi[\delta(\omega + \omega_1 - \omega_{LO}) + \delta(\omega - \omega_1 - \omega_{LO}) \\ &\quad + \delta(\omega + \omega_1 + \omega_{LO}) + \delta(\omega - \omega_1 + \omega_{LO})] \\ &\quad + \frac{1}{2}[N(\omega - \omega_{LO}) + N(\omega + \omega_{LO})]. \end{aligned} \quad (13.9)$$

We see that by the mixing operation, spectral lines will occur at $\pm(\omega_1 + \omega_{LO})$ and $\pm(\omega_1 - \omega_{LO})$. Also, the spectral representation of noise $N(\omega)$ is shifted with respect

to $\pm\omega_{LO}$. We now put the difference to the target intermediate frequency (IF) as

$$\begin{aligned} \omega_1 - \omega_{LO} &= \omega_{\text{IF}} \quad \text{and} \\ \omega_1 + \omega_{LO} &= 2\omega_1 - \omega_{\text{IF}}. \end{aligned} \quad (13.10)$$

This yields the final expression

$$\begin{aligned} S_{\text{mix}}(\omega) &= \frac{a}{2}\pi[\delta(\omega + \omega_{\text{IF}}) + \delta(\omega - 2\omega_1 + \omega_{\text{IF}}) \\ &\quad + \delta(\omega + 2\omega_1 - \omega_{\text{IF}}) + \delta(\omega - \omega_{\text{IF}})] \\ &\quad + \frac{1}{2}[N(\omega - \omega_1 + \omega_{\text{IF}}) + N(\omega + \omega_1 - \omega_{\text{IF}})]. \end{aligned} \quad (13.11)$$

Following the terminology in RF technology [13.16], the frequencies in (13.11) which are not identical to the carrier frequency ω_1 are called the image frequencies, that is, the frequencies $\pm(2\omega_1 - \omega_{\text{IF}})$ and $\pm\omega_{\text{IF}}$. The final step is a dual bandpass operation applied on the IF center frequencies $\pm\omega_{\text{IF}}$.

For simplification, an ideal IF filter, sometimes called a *brick-wall* filter, with a rectangular transfer-function $H_{\text{IF}}(\omega - \omega_{\text{IF}}) = H_{\text{IF}}(\omega_{\text{IF}} - \omega)$ can be applied. This will remove the high-frequency term containing $2\omega_1 - \omega_{\text{IF}}$, where B_{IF} is the single-sided bandwidth of the bandpass filter

$$H_{\text{IF}}(|\omega - \omega_{\text{IF}}|) = \begin{cases} 1 & \text{if } |\omega_{\text{IF}} - 2\pi B_{\text{IF}}| \leq |\omega| \\ & \text{and } |\omega| \leq |\omega_{\text{IF}} + 2\pi B_{\text{IF}}|, \\ 0 & \text{otherwise.} \end{cases} \quad (13.12)$$

The idealized result of the bandpass filter operation for the signal on the IF is

$$\begin{aligned} S_{\text{IF}}(\omega) &\approx H_{\text{IF}}(\omega - \omega_{\text{IF}}) S_{\text{mix}}(\omega) \\ &= \frac{a}{2}\pi[\delta(\omega + \omega_{\text{IF}}) + \delta(\omega - \omega_{\text{IF}})] \\ &\quad + \frac{1}{2}H_{\text{IF}}(\omega - \omega_{\text{IF}}) \\ &\quad \times [N(\omega - \omega_1 + \omega_{\text{IF}}) + N(\omega + \omega_1 - \omega_{\text{IF}})]. \end{aligned} \quad (13.13)$$

We see in (13.13) the two remaining spectral lines of the cosine function at $\pm\omega_{\text{IF}}$. Additionally, the white-noise signals in the pass-band will remain in the IF signal, like it is to be expected. The real situation of the front end is much more complex: as described in [13.9], the LNA and mixer have nonlinear transfer characteristics. Thus, they could potentially generate additional harmonics especially in the case of interference. Let's assume for

a moment that we have spectral lines not only at RF frequency $\pm\omega_1$, that is, assume $\omega_1 = \omega_{IF}$ but also on the IF frequency $\pm\omega_{IF}$ for some reason: we see from (13.11) that in this case, the term $2\omega_1 - \omega_{IF}$ equals ω_{IF} . This means that these components would map to the IF filter pass-band. The same holds for so-called image noise and the oscillator feed-through [13.9]. Band-stop pre-filtering and careful design of the frequency plan are necessary to avoid additional spectral components in the IF pass-band.

Down-Conversion Schemes

In general, two basic approaches exist for frequency handling in the front end: homodyne and heterodyne down conversions. The homodyne approach is also known as the direct-conversion receiver (DCR) or the zero IF receiver. In the superheterodyne approach, the initial carrier frequency is converted to an intermediate frequency before digitization.

Homodyne Down Conversion. In the direct digitization ADC, amplification, sampling, and processing are done at RF. This concept leads to the most simple front end. No mixers and IF filters are necessary. Because the number of analog components is minimized, problems with respect to temperature variations and aging are less. The draw-back is the high power consumption of the ADC and a high computational load on the digital signal processor (DSP). Although this solution is technically feasible for GNSS receivers, leading-edge digital components are necessary, which implicate a high-cost and high-power consuming receiver. However, it is an opportunity for future software-defined GNSS receivers which need multifrequency and multi-band. RF ADCs, for example, from Texas Instruments are able to sample RF frequencies up to 2.5 GHz with 12 Bit resolution [13.31] and a power consumption of 2.2 W per channel.

Superheterodyne Down Conversion. In the superheterodyne down conversion, the amplification and filtering are done at the IF level. It can be done in a single step, which results in a less number of analog parts in the front end. Usually, the digitization is done at an IF of about 4–200 MHz. It provides the advantages of a complete digital solution but relaxes the requirements for the ADC and the follow-on digital processing units by reduced sampling speed. It is a typical solution for single-chip integration at the low end.

At the high end, superheterodyne concepts work with a dual- or triple-step down-conversion scheme, reducing step-by-step the filter bandwidth and thus obtaining very good spectral selectivity. It provides good interference rejection, that is, out-of-band rejection.

The price to pay is in more analog parts and higher power consumption. Associated problems in multiple frequency devices with aging and temperature-dependent group delay have to be considered.

Additional schemes are called baseband or near-zero down conversion. In this concept, the down conversion goes down to 0–100 kHz in a single step. Amplification and filtering are done at RF frequency which results in higher power consumption. In this case, the advantage is that filters and amplifiers are eliminated at IF. However, a price has to be paid in lower interference rejection. This approach could be a solution for single-chip integration where power consumption is not a driver.

Filters

For bandpass filters in the front end, different technologies are possible. In the RF stage, classical inductor-capacitor (LC; from their electrical component representation in circuit theory) circuits and ceramic filters have been used. SAW filters are used in both the RF and IF stages. Filters are determined by the parameters such as center frequency, pass-band frequency, 3 dB bandwidth, insertion loss, attenuation at maximum and minimum frequency, operating temperature range, impedance, and packaging type.

In LC filters, parallel and serial switching circuits with coils and capacitors in a certain complexity level are employed. The complexity of the analog circuits is designed in such a way that the required transfer function of the filter is obtained.

A ceramic filter is a combination of a ceramic resonator (using a mold of ceramic powder under high temperature) with a capacitive coupling network. The center frequency of the resonator depends on the length and the dielectric constant. Ceramic filters are of small size, low-cost, and have a low insertion loss. One drawback of them is that their minimum frequency is 400 MHz (up to 6 GHz). Another drawback could be temperature stability. Ceramic filters have a Chebyshev-type transfer function.

A SAW filter is, in principle, a mechanical filter concept. It makes use of a piezoelectric substrate as a physical basis. The SAW filter has an input transducer and an output transducer. Between these two transducers, the acoustic surface wave travels and couples with the piezoelectric substrate. Usually, the application frequency goes up to 3 GHz with a bandwidth of several MHz. SAW filters can be implemented on quartz (SiO_2). They exhibit a nearly rectangular transfer function and are well described by a Butterworth filter transfer function of higher order.

The transfer function of these filters determines the correlation loss and the shape of the autocorrelation function in post-correlation signal processing.

Computation of System Noise Temperature

For the sake of completeness, the analytical model for the computation of the system noise temperature T_{sys} and the associated noise power density

$$N_0 = kT_{\text{sys}} \quad \text{with} \quad k = 1.38 \cdot 10^{-23} \text{ J/K} \quad (13.14)$$

in 1 Hz bandwidth is reviewed. More details may be found in [13.9] which references [13.32].

Based on the *Friis* formula [13.32] for noise temperature in a cascaded system of stages, where each stage has its own gain factor G_i and noise figure F_i , we obtain the well-known expression

$$T_{\text{sys}} = T_A + \left(\frac{1}{G_1} - 1 \right) 290 \text{ K} + \frac{(F_2 - 1) 290 \text{ K}}{G_1} + \frac{(F_3 - 1) 290 \text{ K}}{G_1 G_2} + \dots \quad (13.15)$$

for the system temperature T_{sys} at a given antenna temperature T_A . By way of example, a system temperature of

$$\begin{aligned} T_{\text{sys}} &= 130 \text{ K} + (1 - 1) 290 \text{ K} \\ &\quad + \frac{(1.83 - 1) 290 \text{ K}}{1.0} + \frac{(1 - 1) 290 \text{ K}}{30} + \dots \\ &= 370.7 \text{ K} \end{aligned} \quad (13.16)$$

is obtained for the system illustrated as in Fig. 13.11.

We see again that the noise figure of the LNA, in this case $F_2 = 1.83$, contributes significantly to the overall system temperature in addition to the antenna noise temperature.

Group Delay Distortion in Multifrequency Receivers

In multifrequency receivers, additional delay problems arise in the front end on different carrier frequencies. Front-end group delay biases (GDBs) are mainly due to the group delay characteristics of filter devices in the front-end chain. Filter technologies used in the front-end on the RF- and/or IF level exhibit a wide range of

group delay variation in the frequency band of interest from 0.5 ns to some tens of ns. Thus, an adequate filter choice is essential to control the overall group delay characteristics of the receiver.

A code division multiple access (CDMA) system with a single carrier is affected by group delay distortion, but in a constant way for all satellites. This problem is mitigated by solving for the clock error term in the navigation algorithm. Because of only small Doppler shifts (± 5 kHz) relative to carrier frequency and signal bandwidth, only small inter satellite GDBs will result on the same carrier.

A frequency division multiple access (FDMA) system with multiple carriers, for example, like the Russian GLONASS [13.33] is tremendously affected by group delay, because the GDBs for different carrier frequencies and/or signal bandwidths could be completely different.

A constant (frequency independent) group delay is common to all signals/frequencies and is removed together with the receiver clock bias in the navigation algorithm. In the general case, the group delay varies (nonconstantly) for different carrier frequencies and over the signal bandwidth, which passes, for example, a wideband receiver, and introduces GDB errors.

A significant improvement of receiver group delay distortion can potentially be achieved by means of moving the filtering functions from the analog filters to highly stable digital filters. Through the current advances in ADC technology, it is potentially possible to perform the analog–digital (A/D) conversion at high IF and, hence, to decrease the number of analog filtering components in the receiver chain. It is also recommended to use identical components in different channel paths to decrease the impact of hardware variation on the interchannel group delay characteristic of the receiver.

In [13.34], group delay effects were investigated for a NovAtel GPS/GLONASS MiLLenium-G receiver card over the GLONASS L1 frequency band from 1602–1616 MHz. The interfrequency GDB relative to GLONASS frequency channel no. 1 was measured and analyzed (Fig. 13.12). The frequency spacing between the GLONASS satellites is known to be 0.5625 MHz. In the test of this high-end receiver, differential pseu-

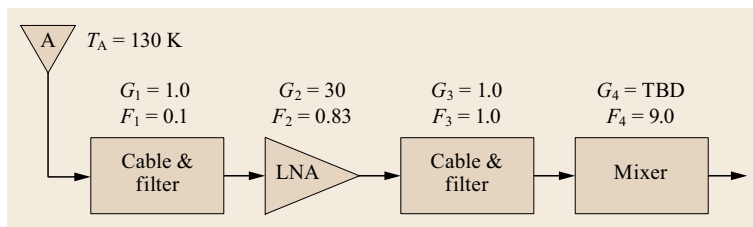


Fig. 13.11 Typical front end of a GNSS receiver adapted for the application of the Friis formula

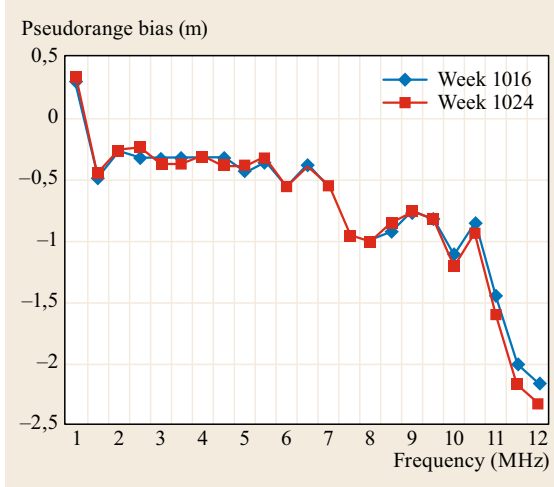


Fig. 13.12 GLONASS pseudorange biases with respect to frequency no.1 over 8 weeks (after [13.34])

dorange biases at the level of $\pm 1\text{--}3\text{ m}$ and differential carrier-phase biases of $\pm 0.02\text{--}0.03$ cycles were found. Aging effects are present with 15 cm per 8 weeks and temperature-dependent effects in the GDB are of the order of $\pm 0.5\text{ m}$ per 10 K change in the operating temperature of the front end.

Group delay distortion is not only a frequency-dependent path delay effect. It also changes slightly the shape (and the inherent timing information) of the chips on the signal. This will finally result in an asymmetrical shape [13.33] of the autocorrelation function after signal processing. From this, it becomes clear that a wide correlator (e.g., with E-L spacing of 1 chip) will be more affected by group delay distortion than a narrow correlator because of different asymmetry levels between 0.1-chip spaced and 1-chip spaced E-L correlation points.

13.2.3 Analog-to-Digital Conversion

In a modern GNSS receiver, the digitization is done immediately after the down conversion, filtering, and amplification. The ADC is a piece-wise linear discon-

tinuous system element in the receiver signal flow. It comprises two basic functions: conversion of continuous time to discrete time (sampling) and conversion of continuous amplitude to discrete amplitude (quantization). Two associated problem areas are related to digitization: quantization noise and/or quantization loss and all related problems to the sampling issue, like aliasing.

The basic elements of the ADC (Fig. 13.13) are a local oscillator to set the sampling frequency, the sample and hold circuit to define the digital time increment of the receiver, a threshold comparator to compare the analog amplitude (on input axis) with a number of n predefined thresholds (Fig. 13.14), and an encoder to convert the result of the comparison (between the amplitude levels) to a certain number of bits (on the output axis).

The number n of thresholds on the input axis could be even or odd. The standard case is that an odd number of thresholds are used. In this case, no level at zero is present and the encoding can be done to 1 bit, 2 bit, \dots , $k\text{-bit}$ words. If n is even, a level at zero (dead-zone) exists. A further decision state becomes necessary (e.g., in the case of two thresholds, that is, $-\Delta$ and $+\Delta$, three levels exist: $-L$, 0 , $+L$). For encoding this logical situation, 1 bit is not sufficient and full 2 bit is not necessary, because we need only three logical states. Because of this hybrid situation [13.12], these quantizers are called 1.5 bit, 2.5 bit, etc.

Based on the classical link-budget for a GNSS channel, it is known that the signal amplitude is at least about 20 dB below the thermal noise level. Therefore, the thresholds of the ADC are dithered by Gaussian noise if no interference is present. The signal itself is deeply buried into noise.

In a digital GNSS receiver, the ADC is a very important subsystem. The implementation used will determine the size of the pre-correlation bandwidth contributions to the implementation loss and will determine the digital data stream, which has to be processed in the signal processing unit (SPU).

In a linear ADC, the transfer function is a straight line with steps around it. It is possible to construct

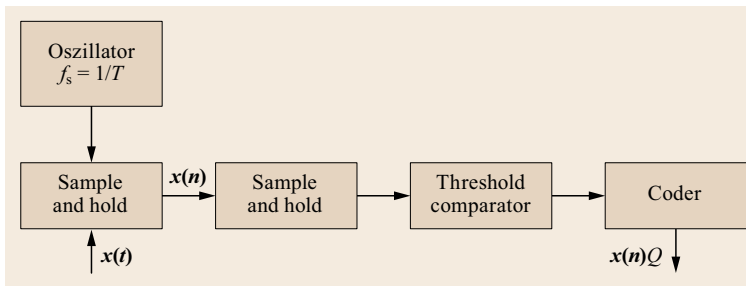


Fig. 13.13 Schematic view of an analog-to-digital converter

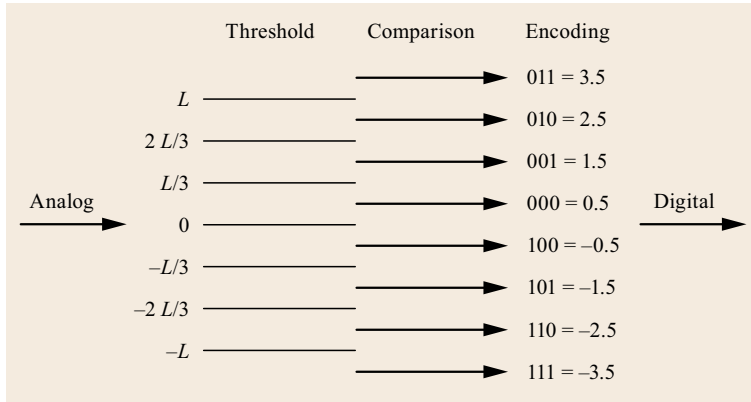


Fig. 13.14 Concept of quantization in the ADC (after [13.9])

nonlinear ADC versions where the location of the thresholds could be changed in an adaptive way and the output level comparison is weighted by applying a weight w .

Sampling Rates

The discussion of sampling rates in a digital GNSS receiver is a more complex issue than it looks at a first glance. Following the sampling theorem, we have to consider in general three cases:

1. Oversampling
2. Nyquist sampling
3. Sub-Nyquist sampling (or under-sampling).

We assume that the front-end filtered signal has a single-sided bandwidth of B in the frequency domain and that the sampling rate is f_s . In digital GNSS receivers, the different types of sampling may be distinguished as follows:

- **Oversampling:** For $f_s > 2B$ in baseband sampling or $f_s > 4B$ in IF sampling, a larger band relative to the GNSS signal band is sampled. This leads to the situation that mainly noise bandwidth and potential interference outside the useful band are sampled. Original white noise is converted to colored noise (correlated noise) which could cause additional issues. If the front end is designed properly, no additional information is gained by oversampling. In this context, a wide-band receiver is not treated as an oversampled receiver.
- **Nyquist sampling:** For $f_s = 2B$ in baseband sampling or $f_s = 4B$ in IF sampling, we sample exactly the useful bandwidth of the GNSS signal. A historical discussion in receiver technology concerns the question whether only the main lobe of a signal should be sampled (which leads to a narrow-band receiver with a standard correlator) or if the entire signal including the side-lobes should be sampled

in the transmitted bandwidth (20 MHz for GPS C/A-code). The latter case leads to a wide-band receiver where narrow correlation is possible. In an IF-sampling receiver [13.9], the I and Q baseband components are obtained directly in the sampling process without mixing with sine and cosine references. As described in [13.9], signals (the I 's and Q 's) are sampled at successive 90° phase shifts. This requires an additional factor of 2 in the Nyquist rate in comparison to sample a single sine or cosine component.

- **Under sampling:** This case is encountered for $f_s < 2B$ in baseband sampling or $f_s < 4B$ in IF sampling. It is clear from the sampling theorem that if a signal has to be reconstructed, Nyquist sampling is necessary. Otherwise, aliasing issues will result. However, it was discovered during the implementation of early software receivers (where processing of very fast digital data stream posed a problem) that in fact no stringent requirement exists to reconstruct the GNSS signal [13.35, 36]. The main requirement on the sampled signal is that it is useable for the computation of an autocorrelation function in the time domain or for a convolution operation in the frequency domain. This means that (against common thinking) under-sampling will work in an autocorrelation receiver. Of course, a price has to be paid which is the aliasing of noise and the reduction of the effective (C/N_0) factor.

We conclude this section with a comparison of baseband sampling and IF sampling. By the Nyquist theorem, a bandlimited signal can be perfectly sampled from a countable sequence of samples if the bandlimited signal is sampled with a sampling rate greater than twice the bandwidth. This fact can also be interpreted that a modulated signal with a null-to-null bandwidth of $2B$ centered at an IF f_{IF} can be perfectly sampled by any sampling rates which necessarily satisfy the Nyquist

theorem (i.e., $f_s \geq 4B$, not $f_s \geq 2(f_{IF} + B)$) without any loss of information, thereby the resulting final IF \bar{f}_{IF} of the aliasing spectrum becoming lower than the original IF (this is a type of down conversion by intentional aliasing). In other words, because the bandpass signal is repeated at integer multiples of the sampling frequency, we can obtain down-converted sampled signal by selecting the appropriate sampling rate. For example, a GPS C/A-code signal ($2B = 2\text{ MHz}$) modulated at $f_{IF} = 95\text{ MHz}$ can be effectively sampled with $f_s = 4\text{ MHz}$ which is extremely undersampling of the harmonic function at IF, and the final IF becomes $\bar{f}_{IF} = 1\text{ MHz}$. This is known as the bandpass sampling technique (also known as the direct conversion or effective intentional aliasing technique) [13.37–39]. In addition to the constraint of the sampling rate by the Nyquist theorem, to avoid unintended aliasing effect, we have to select both the sampling rate and the resulting final IF, not to overlap the information spectrum with the aliasing spectrum [13.40].

Quantization Loss

In the quantizer, the amplitude of the analog signal is converted to a binary representation of the signal (analog-in and binary-out). Depending on the quantization step, q , a certain loss in the signal-to-noise ratio will happen, because the continuous curve is finally represented by a straight line plus a number of rectangular steps. The mathematical description of quantization loss may be found in different sources including [13.9, 41–44]. More or less all of these publications employ a statistical model making use of the probability distribution function (pdf) of a binary signal (± 1) plus Gaussian white noise. We follow the description of [13.41]. The ratio δ between output and input (C/N_0) (which refers to the ADC transfer function) may be defined as follows

$$\delta = \frac{(C/N_0)_{\text{out}}}{(C/N_0)_{\text{in}}} = \frac{\sigma^2}{c^2} \frac{(\bar{N}_S - \bar{N}_{\text{noise}})^2}{E(N_{\text{noise}}^2)}. \quad (13.17)$$

The quantizer loss L may be defined [13.41] as

$$L = -10 \log_{10} \delta, \quad (13.18)$$

where σ^2 is the variance of the additive zero-mean white Gaussian noise (AWGN) process. In this context, only a Gaussian pdf, $N(0, \sigma^2)$, is considered. The quantity c describes the signal content ($+c, -c$), \bar{N}_S is a measure for the square-root of the signal power after quantization, correlation, and accumulation, \bar{N}_{noise} is a measure for the square-root of the noise power after quantization, correlation, and accumulation, and $E()$ is the statistical expectation operator. In order to simplify

the discussion, it is assumed that the signal is on zero IF frequency.

The input signal is sampled in amplitude [13.41] by the ADC through segmenting the input into a series of linearly spaced levels defined as t_i ($i \in -n, \dots, +n$) as illustrated in Fig. 13.15. In this context, t is not the time. The output levels are shown as monotonically increasing. In general, nonlinear assignments of the input range to output state can also be supported. There is an extended region around the zero input levels which yields a zero output level. The weights, w_i , of the output levels are assumed to be 1, 2, 3, etc. as set by most ADCs (linear weighting). But this is not necessarily required as the output levels can be assigned any arbitrary output weight w_i for $t_i < x \leq t_{i+1}$.

In the presence of a signal element c , the conditional probability that the ADC input (x) in the range $t_i < x \leq t_{i+1}$ is occupied is given by

$$P(t_i \leq x < t_{i+1} | c) = \int_{t_i}^{t_{i+1}} P(x|c) dx, \quad (13.19)$$

where $t_{m+1} = \infty$ and $t_{-(m+1)} = -\infty$.

Then, for the binary signal s , which can attain the values $s = +c$ or $s = -c$, the average number in the accumulator after n -independent experiments is

$$E(N_S) = \bar{N}_S = n \sum_i \left\{ \begin{array}{l} w_i P_i(s = +c) p_+ \\ -w_i P_i(s = -c) p_- \end{array} \right\}. \quad (13.20)$$

The terms, p_+ and p_- , are the a priori probabilities that the signal is either $+1$ or -1 (equal to $1/2$ for ideal random sequences). The minus sign associated with the term $P_i(s = -c)$ reflects multiplication by the code replica (de-spreading). The measure of signal power is \bar{N}_S^2 .

The same equation can be used [13.41] to establish the average number in the accumulator when only noise is present. For this case, the probabilities, P_i , are iden-

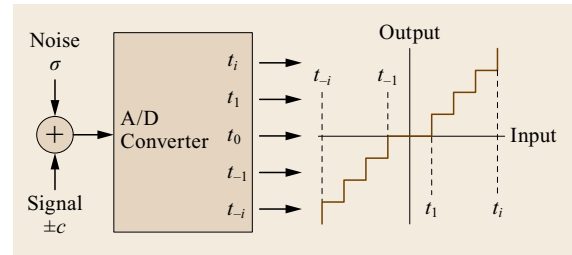


Fig. 13.15 ADC signal and noise model (after [13.41])

tical as $c = 0$, then [13.35]

$$\begin{aligned} E(N_{\text{noise}}) &= \bar{N}_{\text{noise}} \\ &= n \sum_i \left\{ \begin{array}{l} w_i P_i(s=0) p_+ \\ -w_i P_i(s=0) p_- \end{array} \right\} \\ &= 0. \end{aligned} \quad (13.21)$$

The variance of the accumulator number after n independent experiments represents the post-correlation noise variance (note the use of the square of the values in the weight vector) [13.36]

$$\begin{aligned} E(N_{\text{noise}}^2) &= n \sum_i \left\{ \begin{array}{l} w_i^2 P_i(s=0) p_+ \\ +w_i^2 P_i(s=0) p_- \end{array} \right\} \\ &= 2n \sum_i \{w_i^2 P_i(s=0)\}. \end{aligned} \quad (13.22)$$

Based on these expressions, the quantization loss can be computed and depicted in Fig. 13.16.

It shall be outlined that the curves are only valid for infinite (very high) sampling rates. Thus, the results are too optimistic in comparison to the case of a finite sampling rate [13.9]. In [13.9], similar curves are shown for a narrow-band and a wide-band receiver. All curves for the ADC losses approach the two-level (1 bit) case, that is, a loss of 1.961 dB, at low values of normalized threshold Δ/σ . For high normalized thresholds, the loss curves for ADCs which have a level at zero grow without bound, whereas the ADC loss curves stay bounded

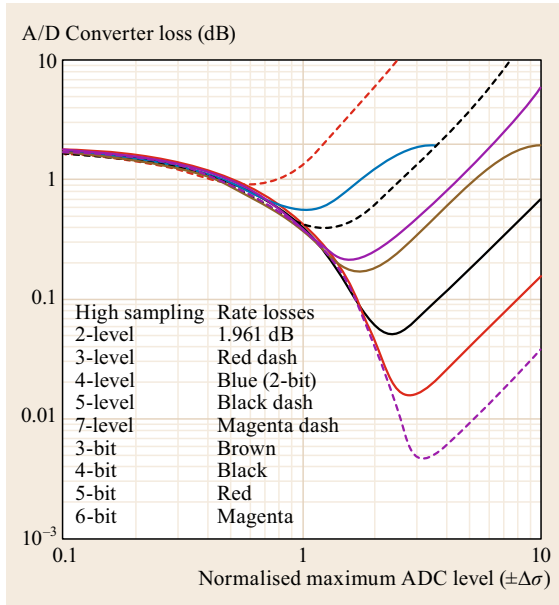


Fig. 13.16 Losses in A/D converters with linear weighting (after [13.41])

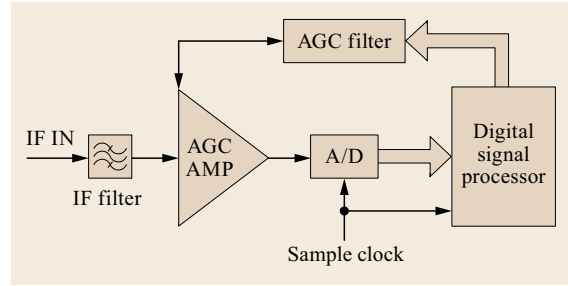


Fig. 13.17 AGC control circuit (after [13.12])

for the case that no level at zero is present, that is, 1 bit, 2 bit, 3 bit.

Automatic Gain Control

In receivers with multibit ADC implementation, the dynamic range of the signal has to be controlled. The goal is that the signal amplitude is conditioned in such a way that the amplitude range falls on average between maximum and minimum thresholds of the ADC. This signal conditioning is obtained by the integration of an amplifier with a variable gain, the so-called automatic gain control (AGC), into the front end (Fig. 13.17). The AGC is driven in a closed-loop feedback control circuit. Some type of measure is needed in the digital domain after the ADC. The measure may be processed when feedback steered the AGC amplifier. Usually, the root-mean-square noise power amplitude is measured and feedback. Monitored AGC levels can be used to detect interference on the GNSS signal.

13.2.4 Oscillators

A GNSS receiver tracks the carrier and code phase of the received signal in reference to the receiver's local oscillator. If either the satellite oscillator or the receiver oscillator shows excessive phase noise, the signal-processing performance of the PLL and DLL with very low cost performance crystal oscillators (CXOs) may be degraded. Concerning phase noise, the PLL is the most critical element, because a replica of the carrier frequency has to be stabilized in the first RF stages, which is at least a factor of 100 larger than the code-chipping rate in common GNSS signals. For some receiver operations which require a long-term clock stability, atomic clock technology on chip scale is available since 2008. A miniaturized cesium oscillator called the chip scale atomic clock was developed by Symmetricom [13.45].

The basic types of oscillators used in today's GNSS receivers and their key properties are briefly discussed in this section. For a more comprehensive discussion of time and frequency standards, the readers are referred to Chap. 5 of this Handbook.

Crystal Oscillators

CXOs use a piece of deformable quartz to generate a precise mechanical oscillation which, at the same time, is converted into an electrical signal. CXOs are built on the principle of the piezoelectric effect. The electrical field interacts with the quartz element and induces mechanical oscillations which can be measured by a pick-off. Because of the mechanical oscillation, a CXO can also be understood as a micromechanical proof mass, which is sensitive to g -load and linear vibration spectra.

Several design versions of CXOs are available (Fig. 13.18). The simplest devices are the low-end crystal oscillators which are used in wrist watches. They provide stability down to 10^{-5} over 1 s. Very important for GNSS receivers are the temperature-compensated crystal oscillators (TCXOs). TCXOs can be produced with the help of a low-cost industrial process. Because their temperature-dependent drift is corrected by a calibration model on the software level and the available performance is between 10^{-6} and 10^{-8} which is sufficient for GNSS receiver signal processing and local clock generation. Higher performance CXOs are also available. In order to get rid of the temperature-dependent drift, the approach is here to install the crystal in an oven. This will provide a very well defined thermal environment. Such crystals are called ovenized crystal oscillators (OCXOs). On the short-term (1 s) scale, these OCXOs could provide stabilities like rubidium clocks down to 10^{-13} . However, cost, linear dimension, and power consumption prevent the use of OCXOs in commercial GNSS receivers. They are used in special applications, for example, in GNSS reference stations to provide a more precise standalone time and frequency reference.

The relative stability of a frequency source is characterized by [13.46]

$$y = \frac{\delta f}{f}, \quad (13.23)$$

where y is the fractional frequency stability, δf is frequency error or jitters, and f is the nominal frequency to be generated. Frequency stability is not a static mea-

sure, but it depends on the application time interval (short term versus long term) on the temperature during operation and on the applied acceleration (g -load).

The temperature-dependent frequency error may be described by a polynomial

$$\frac{\delta f}{f} = a + b(T - T_0) + c(T - T_0)^2, \quad (13.24)$$

with T denoting the actual temperature, a the bias error, b the linear temperature-dependent term, c the quadratic temperature-dependent term, and T_0 the reference temperature, for example, for a laboratory calibration. A typical value for temperature sensitivity of a CXO could have a magnitude of $b = 10^{-5}$ – 10^{-6} K^{-1} for a specified range of $-40^\circ\text{C} \leq T \leq 80^\circ\text{C}$.

Another error term, called the g -sensitivity of the crystal, is dependent on the applied acceleration along a sensitive axis of the CXO [13.47]. The error equation is of the form

$$\frac{\delta f}{f} = K \frac{a}{g}. \quad (13.25)$$

The K -factor describes the sensitivity to acceleration a and g is the gravity acceleration at the Earth surface which is used for normalization in this equation. CXOs for low dynamics ($K = 10^{-8}$) and high dynamics ($K = 10^{-10}$) environments are available [13.47].

Besides these systematic errors, a CXO also shows significant stochastic errors [13.46] which could impair the signal processing in a GNSS receiver.

The consequence of this oscillator phase and frequency noise is that, in a more detailed analysis, the clock error is not really a bias, but a filtered stochastic process and has to be considered with specific dynamics and statistical properties. As a rule of thumb, it is assumed that the clock problem is solved by estimating the clock bias in the absolute or differential navigation solution. This is not really the case, when working with low-cost performance crystals, performing applications in high dynamics and heavy vibration environments, and doing high-precision carrier-phase processing by use of surveying-type receivers. The effect of oscillator phase noise leads to PLL-tracking errors, and if it gets excessive, even to cycle-slips and loss-of-lock.

The variance of frequency stability y , the so-called Allan variance [13.46], is given by the expression

$$\sigma_y^2(\tau) = \text{Var} \left(\frac{\delta f}{f} \right)_\tau, \quad (13.26)$$

where τ is the time constant for performing an average of y by low-pass filtering or integration.

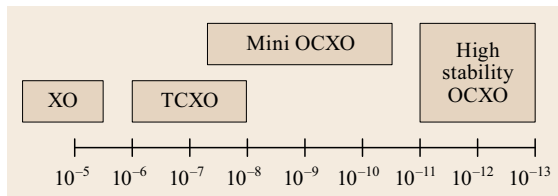


Fig. 13.18 Overview on crystal oscillator technologies for the short term (< 1 s)

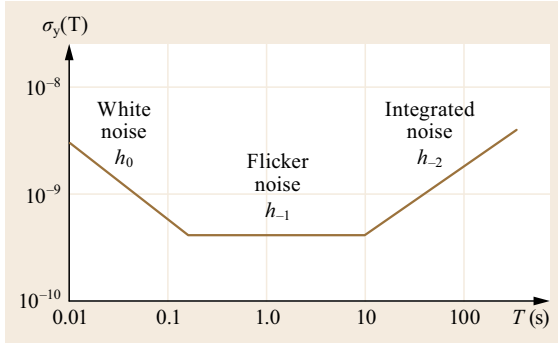


Fig. 13.19 Allan variance pattern as a function of averaging time

Table 13.1 Typical Allan variance parameters for different frequency sources (after [13.48])

Frequency source	White noise h_0 (s)	Flicker noise h_{-1} (-)	Integrated noise h_{-2} (s ⁻¹)
TCXO	10^{-21}	10^{-20}	$2 \cdot 10^{-20}$
OCXO	$2.5 \cdot 10^{-26}$	$2.5 \cdot 10^{-23}$	$2.5 \cdot 10^{-22}$
Rubidium	10^{-23}	10^{-22}	$1.3 \cdot 10^{-26}$
Cesium	$2 \cdot 10^{-20}$	$7 \cdot 10^{-23}$	$4 \cdot 10^{-29}$

In a more explicit form, the Allan variance is given as

$$\sigma_y^2(\tau) = \frac{h_0}{2\tau} + 2 \ln(2) h_{-1} + \frac{2\pi^2}{3} \tau h_{-2}, \quad (13.27)$$

where h_0 , h_{-1} , and h_{-2} represent individual noise contributions that dominate the stability of a frequency source at specific time scales (Fig. 13.19). Typical values of the Allan variance parameters for common oscillator types are collated in Table 13.1.

The power spectral density of the Allan variance, which is of importance in the following analysis, is commonly represented by [13.49]

$$S_y(\omega) = h_0 + \frac{h_{-1}}{\omega} + \frac{h_{-2}}{\omega^2} \quad (13.28)$$

or

$$S_y(\omega) = 4\pi^2 f_0^2 \left(h_0 + \frac{h_{-1}}{\omega} + \frac{h_{-2}}{\omega^2} \right). \quad (13.29)$$

The latter expression already considers the fact that we have to synthesize a reference frequency f_0 in a GNSS receiver for the carrier and code. In order to arrive at the carrier and code level, the frequency expression has to

be integrated

$$S_\phi(\omega) = \frac{S_y(\omega)}{\omega^2}. \quad (13.30)$$

Because the GNSS user is interested in the 1σ phase error in radians, the impact of the loop filter has to be considered

$$\sigma_\phi^2 = \frac{1}{2\pi} \int_0^\infty S_\phi(\omega) |1 - H(\omega)|^2 d\omega, \quad (13.31)$$

where $H(\omega)$ is the transfer function of the loop. This means, the Allan variance power spectral density contributes only to the phase-tracking errors, which are outside the passband of the loop filter (ideal filter assumed). In the case of a second-order loop, that is,

$$|1 - H(\omega)|^2 = \frac{\omega^4}{\omega_L^4 + \omega^4}, \quad (13.32)$$

the phase-tracking error caused by Allan noise in the steady state is obtained as

$$\begin{aligned} \sigma_\phi^2 = & 2\pi f_0^2 h_0 \int_0^\infty \frac{\omega^2}{\omega_L^4 + \omega^4} d\omega \\ & + 2\pi f_0^2 h_{-1} \int_0^\infty \frac{\omega}{\omega_L^4 + \omega^4} d\omega \\ & + 2\pi f_0^2 h_{-2} \int_0^\infty \frac{1}{\omega_L^4 + \omega^4} d\omega, \end{aligned} \quad (13.33)$$

where $\omega \approx 1.9B_L$ (at a loop-damping factor $\xi = 0.707$) and B_L is the single-sided noise bandwidth of the loop. Evaluation of the integrals leads to the following expression [13.49]

$$\sigma_\phi^2 = 2\pi^2 f_0^2 \left(\frac{\pi^2 h_{-2}}{\sqrt{2} \omega_L^3} + \frac{\pi h_{-1}}{4 \omega_L^2} + \frac{h_0}{4 \sqrt{2} \omega_L} \right). \quad (13.34)$$

The second effect is the induced phase noise on the CXO caused by vibrations of the user platform. The final phase noise-tracking error depends on the g-sensitivity of the oscillator and the power spectral density of the vibrations. For this case, the expression is of the form

$$\sigma_\phi^2 = 2\pi f_0^2 k_g^2 \int_0^\infty G_g(\omega) \frac{\omega^2}{\omega_L^4 + \omega^4} d\omega, \quad (13.35)$$

where k_g denotes the g-sensitivity of the oscillator and $G_g(\omega)$ is the single-sided vibration power spectral density (measured in units of $g^2/(\text{rad/s})$). Representative

g-sensitivity values are given by

$$k_g = \begin{cases} 1 \cdot 10^{-9} \text{ g}^{-1} & \text{low dynamics receiver,} \\ 3 \cdot 10^{-10} \text{ g}^{-1} & \text{high dynamics receiver.} \end{cases} \quad (13.36)$$

In this case, an analysis is difficult to perform, because the result depends heavily on the vibration spectrum under which the user has to operate on a specific platform or vehicle. In harsh dynamic environments, the tracking error caused by vibrations can become significant, again with the consequence that loss of lock can occur.

Chip-Scale Atomic Clock (CSAC)

Besides crystal oscillators chip-scale atomic clocks (CSACs) are commercially available. The CSAC technology was developed in a Defense Advanced Research Projects Agency (DARPA) program in several project phases in 2001. The production started in 2008. The highly portable atomic clock consists of a physics package including the cesium cell, shielding, and a lid. The system elements are integrated on a printed circuit board (PCB). Currently [13.45], the size is about $4.0 \times 3.6 \times 1.3 \text{ cm}^3$, the weight is 0.035 kg, and the power consumption is 120 mW while in operation. The cesium atoms are excited by a laser beam of a vertical-cavity surface-emitting laser (VCSEL). The laser output signal (the resonant line) is detected by a photodiode. The stability of the clock in terms of Allan deviation is $1.5 \cdot 10^{-10}$ over 1 s and $5 \cdot 10^{-12}$ over 1000 s which outperforms a typical TCXO. An additional advantage is the low sensitivity with respect to acceleration and vibration levels in comparison to a CXO. The temperature sensitivity is less than $5 \cdot 10^{-10}$ over the entire operating temperature range (-10°C to $+70^\circ\text{C}$). A low (about 10^{-10}) sensitivity concerning input voltage variations and applied magnetic fields can be observed. CSAC technology will evolve to a smaller size, lower power consumption, and lower cost. The advantage of CSAC technology as a frequency source for GNSS receivers is due to the ability to acquire or re-acquire even longer codes with a short time-to-first-fix (TTFF) based on a very stable time scale. Because of lower phase noise in the tracking loops, it allows for longer integration times (coherent integrations) in signal processing. Because of its precision, a CSAC can be used for clock-coasting to substitute for the clock-error term in the estimation filter. A lower number of visible satellites are required in such a case.

13.2.5 Chip Technologies

In miniaturization and integration of GNSS receivers, the state of chip technologies plays an important role. Mainly the form factor, the power consumption, and

production cost of a receiver depend on the current state of semiconductor technology and the semiconductor processes which can be applied. In a GNSS receiver, chip technologies are used in the RF domain, in the SPU, and in the navigation processing unit (NPU). Additionally, memory units, like read-only memory (ROM) and random access memory (RAM), and (if applicable) cryptographic units depend on microchip designs. The degree of very large scale integration (VLSI) decides on the use of the GNSS function in mass market devices, that is, in mobile phones, personal digital assistants (PDAs), automotive systems.

The development trend in VLSI GNSS chips is worthwhile to consider: even in the GNSS, mass market domain chip technology is lagging behind the leading-edge semiconductor development by 2–3 innovation cycles or technology nodes. The reason is that the market size of GNSS receivers is significantly lower than the market size of systems like PCs, mobile computing and communication devices. According to Intel, one billion devices like notebooks were used in the portable PC market in 2005. The forecast for 2015 for various ways of ubiquitous computing is above 10 billion units. In comparison to this, about 6 billion mobile communication devices are on the market [13.50]. About 1 billion civil GPS C/A-code chips are in use today (2013). Clearly, the non-GNSS information technology markets drive the development efforts in chip complexity (> 50 million transistors per die) and the minimum feature size down, for example, to 22 nm (Intel Atom) processor [13.50]. Another problem area exists for VLSI chip integration in the area of, for example, military receivers which have to fulfill higher military specifications (MIL-SPECS). Because with the decreasing feature size, only a few international foundries are able to produce the leading edge SiO_2 -integrated circuits. The investment in these foundries is getting more and more expensive, for example, because of more advanced lithographic tools [13.51] and larger diameter wafer-processing facilities. It gets difficult for semiconductor companies in the defense sector to keep their pace. This has led to the interesting situation that in terms of high integration mass market products are outperforming silicon integration in the military receivers.

The basic functional silicon element is called a die (or sometimes called the bare chip). Because of packaging and pins, a die is only a fraction of a chip. It is diced out of the wafer. On a chip, there could be many dies with different functions. Sometimes such systems are called systems-on-a-chip (SoC).

Digital Chip Technologies

The processing requirement (based on experience) of a six-channel GPS C/A-code receiver is of the or-

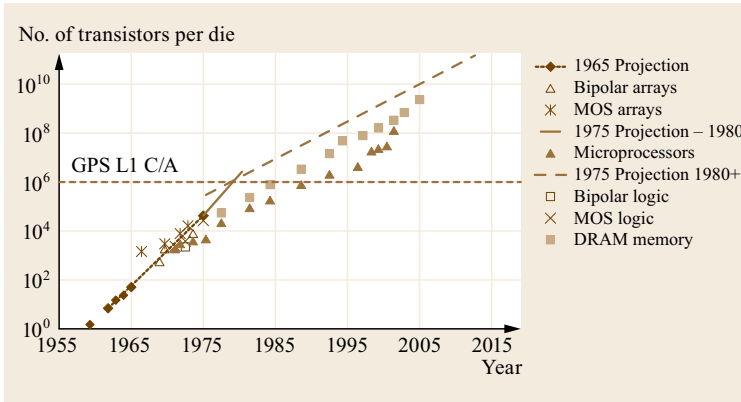


Fig. 13.20 Complexity (number of transistors per die) of MP and memory (after [13.51], courtesy of the Institute of Electrical and Electronics Engineers (IEEE))

der of 80–100 MIPS. In information technology, it is questioned if the parameters MIPS (mega instructions per second) for measuring processing power are adequate. However, we will use it here to define orders of magnitude. For a processing power of 100 MIPS, an equivalent of one million transistors on a SiO_2 die are needed. Interestingly, these values hold for an Intel 486 up to an Intel Pentium PC processor. Since these processor generations were available, it was possible to implement the signal processing for the C/A-code on a PC (software receiver).

The dramatic progress of semiconductor technology in digital chips is always referred to *Moore's law*. Gordon E. Moore was a pioneer in early integrated circuit development working for Fairchild Semiconductors. In 1968, he had been a co-founder of Intel. He observed that every 2 years the number of components on a silicon area doubled (Fig. 13.20) and that the overall processor power increases in a similar manner (Fig. 13.21). In 1965, he published his work known as *Moore's Law* and gave a forecast over the next 10 years. Around 1975 [13.51], the rate of integration complexity slowed down. The reason was that in the first phase it was tried to make use of wasted area on the silicon chip. This possibility was exhausted around the mid-1970s. It is clear that Moore's law is an empirical rule. It was gained by several factors: leading manufactures like Intel tried to keep close to it, competitive manufactures did the same, and the semiconductor design tools were developed with respect to Moore's target.

Since a decade with a discussion has taken place about when *Moore's law* will finally end. There is evidence that it will end, because circuits cannot get smaller than atoms, lithographic techniques become more and more critical because the structures get smaller than the wavelength of light, quantum effects get more and more pronounced with finer structures. However, several opportunities exist [13.52] to shift the limit: diversification (power and polymer electronics) and performance en-

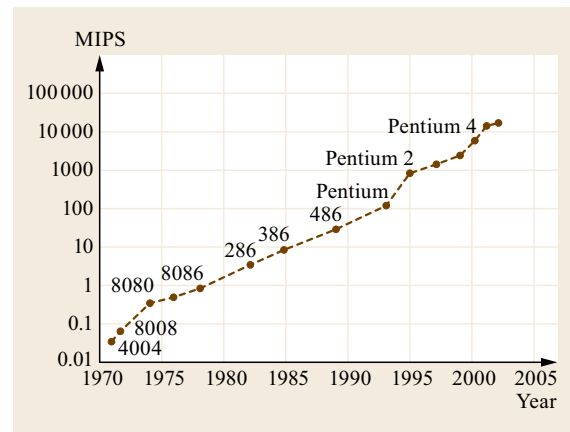


Fig. 13.21 Growing processor performance (after [13.51], courtesy of IEEE)

hancement (multicore processing, photonics, graphene transistors).

Based on the discussion of Intel [13.33] and the semiconductor community, the prediction is the following: Moore's law will continue up to 2015+ and the size of structures will be reduced by a factor of 2 every 6 years (Fig. 13.22).

The width of smallest structures achieved was 65 nm in the year 2005 and 22 nm in the year 2011. The latter is 2013 in production for PC processors. Research [13.33] for the upcoming years 2015+ will be in the area of 10 nm, 7 nm, 5 nm where lithography, interconnections, materials, and other improvement areas are investigated.

A further element of Moore's law is that the semiconductor complexity (number of transistors) will increase by a factor of 4 every 6 years, processing performance increase by a factor of 150 at the same power consumption level until 2015+. The cost per transistor is halved every 2 years.

The same processing performance at decreasing power consumption will be obtained: for a GNSS dig-

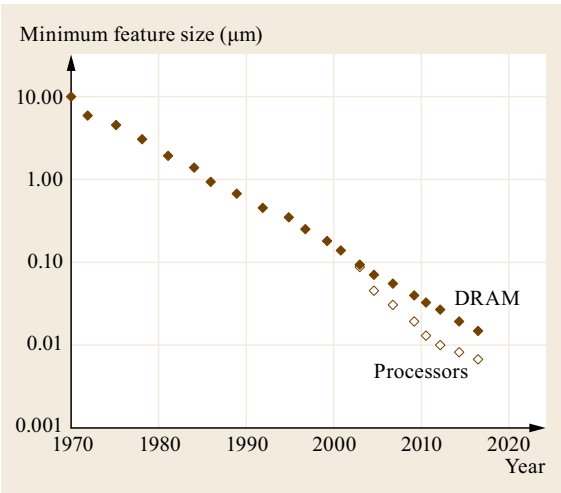


Fig. 13.22 Decreasing minimum feature size in μm (after [13.51], courtesy of IEEE)

ital chip, this could mean that today 100 MIPS will consume 100 mW and in 2015+ 100 MIPS will consume 1 mW only.

RF Chip Technologies
The RF semiconductor technology and its future development is a driver for the front-end architectures. Moreover, many new integrated or low-cost applications became only possible by means of RF chips. The current development tendency is toward lower power consumption, while integrating RF parts and enabling multifunctional operations.

As outlined before, digital electronics is governed by Moore’s law. In the digital domain, the integration of a high number of transistors on a chip is important. Electrically, the silicon (tri-gate) metal oxide field effect transistor (MOSFET) is dominant. In semiconductor-based RF electronics Moores’s law and VLSI are less important. Historically, RF semiconductors were developed till the 1980s for military systems like radars. Based on the hype in the field of mobile communication around 1990, the RF chips market moved clearly for various consumer applications.

The main requirement on an RF transistor is its fast reaction on a change of the input signal [13.53]. The two basic areas to achieve this are transistor design and selection of the suitable semiconductor material. Nowadays, RF transistors are used in a range between 0.5 and 100 GHz. In general, two [13.53] basic transistor concepts have been developed. The field-effect transistor (FET) and the bipolar transistor (BT). In an FET, the output current (drift current) is controlled by an orthogonal field. Conductivity is controlled by the potential of the gate. Three types of FETs are known [13.53]:

metal–semiconductor FET (MESFET), high electron mobility transistor (HEMT), and metal–oxide semiconductor FET (MOSFET). In bipolar transistors, the output current is controlled by voltage across the p–n junction. Two types [13.53] have been developed. The bipolar junction transistor (BJT) and the hetero junction bipolar transistor (HBT). The performance [13.53] of the RF transistor is described by their ability to amplify, gain, frequency limits, output power and minimum noise figure (Table 13.2). For GNSS receivers, the noise figure F of the LNA is important. Additionally, the power consumption of the RF chip and its production costs are relevant.

It should be outlined that the noise figures F presented in the table refer to a single transistor. Because in an LNA, several transistors are integrated, the total noise figure could be higher. The following materials are considered: iridium phosphide (InP), silicon (Si), gallium arsenide (GaAs), and silicon germanium (SiGe). The physical gate length in a microwave silicon ASIC was in the year 2001 on a level of 90 nm and in the year 2016 it will be 11 nm.

Complementary metal oxide semiconductor (CMOS) is the cheapest and widely used silicon semiconductor process to develop digital, that is, logic chips. In CMOS, p-channel and n-channel MOSFETs are used on a common substrate. RF CMOS chips support frequencies up to 3 GHz which is more than sufficient for GNSS LNAs in L-band. For low-end single-chip integrations, RF-CMOS is important because digital functions and RF functions can be integrated within the same semiconductor process.

Digital Signal-Processing Units
The digital SPU is traditionally decomposed into different processors (Fig. 13.23). The reason for this hybrid architecture is that we need, on the one hand, very fast parallel operations to be applied on the fast data stream which is output by the ADC. The ADC output data rate can be on multiple 10 MHz level depending on the bandwidth of the signal, receiver, and the applied sampling rate. On the other hand, we have to process signals of medium speed like the integration and dump data (output of the correlators) which are typically less than

Table 13.2 Minimum noise figures F for different RF transistor materials (after [13.53])

Transistor type	F (dB) at 2 GHz	F (dB) at 5 GHz
InP HBT	0.4	1.1
Si BJT	0.8	1.2
GaAs HBT	0.9	1.2
SiGe HBT	0.1	0.5
GaAs MESFET	0.1	0.2

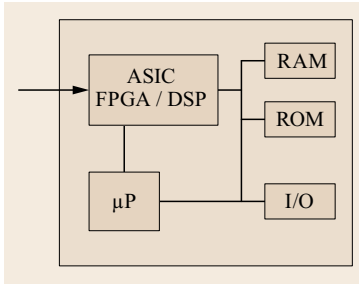


Fig. 13.23
Generic structure
of digital GNSS
receiver section

the kHz level. Additionally, we have signals with low speed such as the tracking loop processing (< 100 Hz) and navigation processing (< 10 Hz). Besides data rate consideration, the complexity of mathematical functions and the necessary instruction set are also issues.

The generic standard partitioning of the SPU is to implement the fast (and more elementary operations) process on a digital signal processor (DSP) (Table 13.3). The DSP could be augmented by one or several embedded field programmable gate arrays (FPGAs). An FPGA can be re-programmed and, thus, allows for a higher flexibility. However, it is expensive and power consumption is significant. Instead of the FPGA, an ASIC can be used, which is the cheapest way to implement the digital signal processing logics but it is only reconfigurable in a limited way. Additionally, an MP or core processing unit (CPU) is always present in the SPU. The MP (μ P) is controlling the entire SPU and the input/output (I/O) functions. In a software receiver, it is tried to comprise all processing functions in software running on a general purpose computer.

A DSP is, in principle, a specialized MP adapted to the typical requirements of fast real-time digital signal processing like digital filtering or fast Fourier transform (FFT) computation. A standard application is the ability to process correctly and with high data throughput *fast multiply and accumulate* (MAC) functions like finite impulse response filters (FIR) with many coefficients α_k

$$y_k = \sum_{k=0}^m \alpha_k x_{-k} \quad (13.37)$$

The providers of DSPs are limited. The big players are Texas Instruments (TI), Analog Devices (ADI), and Freescale (formerly Motorola). A DSP that has been

Table 13.3 Partitioning of GNSS receiver processing (after [13.54])

Processing task	Processing hardware
Autocorrelation	ASIC/FPGA/DSP (CPU)
Acquisition	DSP/CPU
Tracking	DSP/FPGA/CPU
Navigation	CPU (DSP)

used from the beginning on in GNSS receivers is the Texas Instruments TMS320 family [13.55]. Over the years, many generations and variants were available. These DSPs provide a processing power between 60 and 2000 MIPS, make use of a 24 bit fixed point or 32 bit floating-point architecture, and provide standard functions for digital filtering. The TMS320 can be programmed in C, C++, and Assembler.

On the CPU side, some GNSS receivers make use of ARM processor chips. ARM stands for the advanced reduced instruction set (RISC) machines [13.56]. From the beginning, in the early 1980s, eight versions of ARM processors have been developed. ARM version 6 and later provide a clock rate of 750–2000 MHz, a 32/64 bit architecture and a computational performance between 60 and 180 MIPS.

General purpose machines (GM), which have been used for the building of software receivers, belong in many cases to the PC processor family like Intel Core 2 (e.g., use of the Lippert Toucan board, 2 GHz CPU clock rate, 40 W power consumption). Other high-end options are the use of the 3.2 GHz cell broadband engine which is said to have a higher processor speed (factor 30 in comparison to Core 2). The cell broadband engine is used in the PLAYSTATION 3 but is a candidate for future software receivers.

GNSS Receiver Integration Levels

Apart from functional solutions, several architectural concepts can be applied for the receiver implementation (Table 13.4). The architectural design used depends mainly on the target of the development (prototyping, geodetic receiver, chip-set, single chip), on the available investment budget and last but not least on the time schedule for the development. Higher integration has several advantages [13.57]: smaller form factor by reducing the number and size of parts, higher reliability by removing solder points of pins, lower power by removing the resistance of pins and electrical circuits, and reduced production cost in the application of semiconductor manufacturing. Higher integration will improve digital performance because of shorter electrical lines and higher processing speeds. However, in higher integration, interference between the RF and digital section can become an issue.

The individual architectures are briefly described below:

- **Discrete printed circuit board (PCB) design:** This architecture requires the largest area and the highest number of parts but allows for the use of high-performance discrete devices. Having the least initial production costs, it is very suitable for low- or medium-size productions. It is the technology

Table 13.4 Receiver integration levels (after [13.57])

Integration level	Architecture	Technologies	Relative R and D cost	Market
Separate LNA/RF/ASIC/ μ P	Discrete PCB	GaAs, SiGe, Bipolar Si, CMOS	1	High-precision, scientific applications
(LNA+RF)/ASIC/ μ P	Hybrid MMIC or RF Chip PCB	GaAs, SiGe, Bipolar Si, CMOS	1.5–2	High-end, surveying, aviation
(LNA+RF)/(ASIC+ μ P)	Dual chip PCB	Bipolar Si, SiGe, CMOS	3–4	Low-end, handhelds, cars
(LNA+RF+ASIC+ μ P)	Single chip	CMOS	6–10	Ultimate mass market, E-911

- for prototyping and early bread-boarding receivers. LNA and RF filter are usually integrated together in the L-band antenna.
- *Hybrid monolithic microwave integrated circuit (MMIC)*: This can be seen as a merge between the discrete PCB and chip-set design for medium-size production. This approach was used in surveying type of receivers in the 1990s by employing high-performance front ends (multiple step down-converters).
 - *Dual chip design on PCB*: This architecture is based on an integrated RF chip and on an integrated digital chip. This chip-set will result in lower size, power consumption and price at the expense of higher initial costs and, therefore, is only applicable to consumer market applications, where highest integration and extremely low-power consumption is not a driver. Usually, simplified front-end techniques (single-step to baseband) are used.
 - *Receiver single-chip integration*: Integration of the RF unit and the SPU into a single chip results in the cheapest (production cost) and smallest solution with the lowest power consumption. Because of the very high initial cost, for example, for a semiconductor design in 65 nm CMOS, the single-chip solution is only applicable to ultimate mass market. The key for the implementation of a single-chip receiver is that the semiconductor process (low-cost CMOS) for the RF part is the same as for the digital part.
 - *Multifunctional chips*: Sharing the chip area with other functional elements like the communication channel. In future, it could be the dominant technology in mass market applications.

13.2.6 Implementation Issues

From the previous section, it became clear that a GNSS receiver is more or less a high-speed calculating machine. Thus, the classification of a receiver as a hardware receiver or a software-defined receiver has mainly to do with the kind of processing architecture which is used. The signal-processing logics and data flow to

be handled are the same for all the different processing architectures. The signal structure to be processed is defined in the respective interface control documents (ICDs). We will briefly discuss the advantages and disadvantages of the three main processing architectures:

- *Software-defined radio based on a general purpose machine*: The advantages are the use of a fully programmable MP, that is, no hardwired silicon is used. The general purpose machine (GM), like PC-based MP, or DSP is accomplishing all the digital signal and navigation processing and is highly reprogrammable by use of higher programming languages, for example, C/C++. The software receiver was in its early stage a research project pursued by some universities. The rationale for industry to keep this development in focus is that modern portable navigation device (PNDs) will have in future even faster and more powerful CPUs and larger memory. Using a single CPU will help to reduce the number of different chips in PNDs.
The disadvantages of the software receiver are that the processing load on the GM is highly critical, and also the power consumption is high. Because of the GM board, the resulting system is form-factor intensive. Software testing and validation is also a tremendous issue. It seems to be clear that the GM-based software receiver will penetrate the high-end market of hardware GNSS receivers like reference stations because these applications are independent of form factor and power-consumption requirements.
- *Software-defined radio based on FPGA*: The typical baseband processor in this configuration is based on the partitioning of functions on two processing cores: CPU and FPGA. Both elements are integrated on the same PCB and communicate together. Fast operations are done on the FPGA, whereas lower speed operations and system control are done on the CPU.
The advantages here are that re-programmability leads to a configurable receiver architecture and re-

duced development costs because of easy bug fixing by afterward software update capability.

The disadvantages lie in the relatively high production cost (> 2000 US \$ procurement cost for the bare FPGA board), especially for large and multiple FPGAs on a single board. Additionally, special software development tools (e.g., Verilog/VHDL, Handel-C) are necessary. The board size determines the form factor of the receiver. Also, relatively high power consumption depending on the number of CPUs and FPGAs and digital clock rate could result.

- **Hardware-defined based on ASIC:** In a hardware receiver, the code replica generation, the mixing with the trigonometric reference signals (early, punctual, late), and the integration and dump operations for a bank of, for example, 12 channels are conventionally implemented in a hard-wired logic on an ASIC. The advantages of the ASIC approach are low production cost for silicon (CMOS) in very large volumes. The power consumption of highly integrated semiconductors is the lowest option available for GNSS receivers. Large-scale integration leads also to a small form factor.

The major disadvantages are low flexibility because of the fixed and hardwired chip design. If new navigation signals are transmitted and/or signal designs are modified, the chip is not usually adaptable to such changes and the receiver, has to be bought anew. For mass market receivers this is not posing a big issue because the innovation cycle is only several years, for example, in mobile phones. For more expensive hardware-based receivers (like surveying and military receivers), the not reconfigurable ASICs are an issue. Usually, backward compatibil-

ity is expected in the case of new signal structures. A certain way out of the problem is to use a DSP besides the ASIC and to attribute some functions on the DSP. Again for ASIC design, special development tools, for example, hardware development languages (HDLs), are necessary. High development cost due to additional chip runs in the case of bug fixing or architecture changes could result. The development and design costs for a VLSI chip are the highest in comparison to other implementation schemes.

The three basic implementation schemes which were described may be compared with each other in terms of power efficiency and area efficiency (Table 13.5; [13.54]). MOPS means mega operations per second and area refers to the chip area necessary to provide the required complexity for the provision of the processing power envisaged. It is notable that between a general purpose software receiver processor and an ASIC, a disadvantage of 10^4 exists in power efficiency. The area efficiency of an ASIC is by a factor of about 10^5 superior in comparison to a general purpose machine. It follows from this that PC, ASIC, FPGA solutions are justified solutions, but we have to be clearly aware of their advantages and disadvantages. It is a naive thinking that Moore's law will solve all the processing problems: if PC processors get better by using higher integrated CMOS, the same will hold for ASICs. Thus, to assume that a competition exists between software- and hardware-based receivers, in all the market segments is not a very valid assumption. For pure software receivers a growing niche market may be identified at the high end, in prototyping and research applications.

13.3 Multifrequency and Multisystem Receivers

In the former, the subsystems of a GNSS receiver were described in a more generic way. No effort was made, besides the discussion of front-end group delay, to look into architectures which make use of multifrequency and multisystem GNSS. Using more systems and more frequencies will lead to higher complexity: multifrequency antenna and front end, more complex frequency plans, faster ADC data streams, and higher parallel processing power requirements on the SPU because of more parallel channels. In this section, we intend to briefly present the major design characteristics for a GNSS receiver in the context of GPS modernization, Galileo, GLONASS, and BeiDou development.

13.3.1 Civil Receivers for GPS Modernization

Considerations for L2 Civil Signal

With the launch of the GPS Block IIR-M satellites (since 2003), the L2 civil signal (L2CS), which is the second civil GPS frequency in the context of GPS modernization, is transmitted. The detailed design of the signal was initially described in the ICD-GPS-200D. The spectral characteristics of L2CS are not so different from the C/A-code on L1. In general, the L2CS is generated through chip-by-chip multiplexing of a code of medium length (CM: 10.230 chips, period of 20 ms) and a long code (CL: 767.250 chips, period of 1.5 s). Longer codes are used in comparison to the C/A-code

Table 13.5 Power efficiency and area efficiency of different receiver implementations (after [13.54])

Receiver type	Platform	Characteristic	Power efficiency	Area efficiency
Software defined	PC or DSP	Programmable	10 mW/MOPS	1 MOPS/mm ²
Software defined	FPGA	Reconfigurable	0.1 mW/MOPS	100 MOPS/mm ²
Hardware defined	ASIC	Dedicated	0.001 mW/MOPS	10 ⁵ MOPS/mm ²

in order to reduce cross-correlation between different codes. The final chipping rate after multiplexing is 1.023Mc/s as in the L1 C/A-code case.

Thus, an L2CS receiver channel is very similar to an L1 C/A-code receiver. In general, the same front-end concept may be used in terms of filtering and bandwidth as in a C/A-code receiver. However, a synthesizer at $L2 = 1227.6$ MHz and a respective down converter to a reasonable IF (or the same IF as used for L1) is necessary. The main challenge for the implementation of the L2CS lies in the digital domain. The direct acquisition of the long code (CL) seems to be impractical without a precise GPS time estimate of the receiver [13.58]. Therefore, a pre-acquisition of the C/A-code L1 and/or the moderate-length code (CM) L2 with the subsequent handover of timing and frequency information seems to be reasonable. As outlined in [13.58], three implementation options exist:

- CM only for signal tracking and data demodulation
- CL only for signal tracking and CM for data demodulation
- Use of both CM and CL in a combined mode to track CL for the ranging signal and CM for data demodulation.

Another focus could be the implementation of long coherent integration (over 1.5 s or multiples of it) during autocorrelation processing because of the data-less CL channel property. Extremely long coherent integration poses higher requirements on the oscillator phase noise (oscillator quality), because the phase changes during the integration interval. The same holds for signal dynamics (motion of receiver): loop aiding from the CM tracker could be necessary. In principle, the L2CS has the ability to replace the semicodeless P(Y) techniques in high-end receiver integrations after December, 2020. Beyond this date, no further guarantee for the P(Y)-code is given by the US government. The GPS-ICD-200 published the signal design for the P-code, but no guarantee was given for codeless and semicodeless techniques.

Considerations for L5 Frequency

The third civil frequency in the GPS modernization program is the L5 signal which was mainly intended to support safety-of-life applications in aviation and other similar fields. The L5 signal design is described in the

ICD-GPS-705 in detail. It was firstly transmitted with the launch of the Block IIF satellites beginning in May, 2010 and will be transmitted from the GPS III constellation. The signal is transmitted in L-band on $L5 = 1176.45$ MHz in a bandwidth of about 24 MHz and has a chipping rate of 10.23 Mc/s. Some additional features which are important for L5 receiver implementation are that besides the spreading code a secondary Neuman–Hoffman (NH) code is used. The concept is to achieve better cross-correlation properties by increasing slightly the spectral line spacing of the periodic spreading codes. Additionally, a QPSK modulation is basically used, where the Q-channel is again a data-less channel, and, in the I-channel, navigation data (L5 data message) is present. The navigation data is forward error correction (FEC) coded to a rate of 100 sbs (symbols per second). A 24 bit cyclic redundant code (CRC-24) is added to the navigation data.

The main issue comes from the fact that the L5 signal is collocated with other civil and military signals namely distance measurement equipment (DME), tactical air navigation (TACAN) and joint tactical information distribution system (JTIDS) together with the multifunctional information distribution system (MIDS) and near-band radars. These signals cause pulse type interference for the L5 receiver channel. Civil C/A-code receivers are not designed to work in pulse-type interference environments. However, in future GNSS receivers, the capability to mitigate this type of interference is needed. This holds especially for the new aviation frequencies GPS L5 (and Galileo E5) but may pose a problem also for the wide-band Galileo E6 frequency band, where in some European countries, high-power civil and military radars and amateur TV coexist.

Following [13.58], some major changes in the L5 front end with respect to a GPS L1 C/A receiver have to be made: an RF pulse limiter to provide burn-out protection for RF components in the case of strong pulses, an RF bandpass filter of 20 MHz bandwidth on $L5 = 1176.45$ MHz, an additional frequency synthesizer and down converter from L5 to IF, and an analog or digital pulse blanker for the suppression of the pulse interference.

The basic change with respect to a conventional front end is the implementation of a pulse detector and pulse blanker into the signal flow between AGC and ADC in the L5 front end (Fig. 13.24).

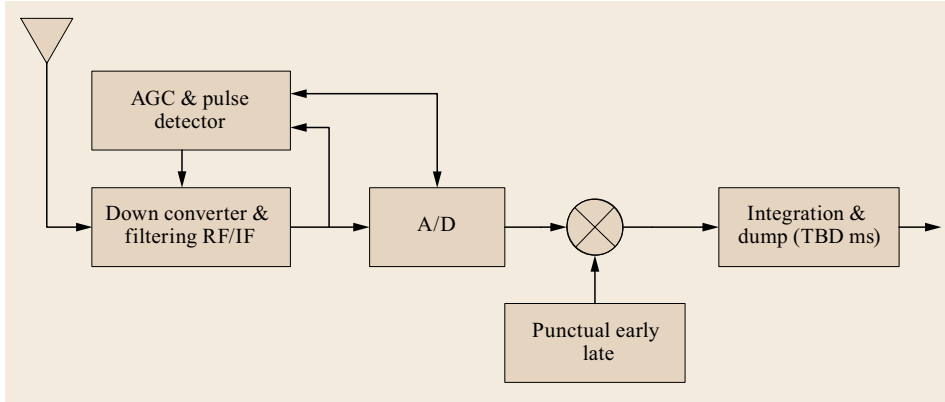


Fig. 13.24 Pulse blanking receiver architecture (Radio Technical Commission for Aeronautics (RTCA) SC-159 L5 concept)

In general, the effects of pulsed interference on the GNSS receiver can be completely different [13.59], dependent on the characteristics of the interfering pulsed signal (peak, power, duty cycle, pulse duration) and on the receiver implementation (technology and design of front end, AGC, single- or multiple-bit ADC, correlation and tracking software). Depending on their peak power level pulsed interference leads to the following problems in a GNSS receiver:

Strong pulses will, on the one hand, saturate the front-end components and, on the other hand, saturate the ADC [13.59]. Typical GNSS receiver front-ends have only a restricted linear region to transfer the antenna power to the quantization result of the ADC. Outside the linear region, the front-end saturates, that is, it is not able to follow higher power variations (will output a constant gain). After a very strong pulses, it could occur that the front end will not recover (or takes long to recover), even if the pulse is zero. Recovery times in the microsecond range are possible for power levels of > 20 dB above saturation. A typical value for a commercial front end is 40 ns/dB [13.59]. The front-end behavior in the case of pulses depends to a large extent on the front-end technology used (bipolar transistors, MMIC, RF chip, etc.).

In addition to front-end problems, strong pulses could saturate the ADC, if the signal level is above the maximum threshold. ADC saturation leads to the positive fact that no high-pulse energy enters into the SPU, but the useful GNSS signal is completely lost.

Weak pulses are those which will not saturate the front end and the ADC [13.59]. However, they cause degradations in the receiver-tracking performance. Weak pulses add more or less noise I_0 to the thermal noise floor. As in the case of narrow- or wide-band interference, the C/N_0 is degraded to

$$(C/N_0)_{\text{eff}} = \frac{C}{N_0 + I_0}, \quad (13.38)$$

where I_0 depends on the power spectral density of the pulsed signal and on the type of correlator. Demodulation of data is also very sensitive [13.59] on the pulse duration, especially for cases of the high symbol rate (small pre-detection integration time relative to pulse duration).

In principle, three solutions to mitigate pulse-type interference are known:

- Pulse clipping: the power above the ADC threshold is removed
- Pulse suppression: the pulse is driven to the noise level with fast ADC and a wide-band front end
- Pulse blanking: pulse is detected in the front end, signal and noise are driven to zero, and ADC generates only zeros.

In any solution, the degradation of the C/N_0 will result. Because blanking suppresses the interfering signal but at the same time completely suppresses the thermal noise, it is superior to the other two methods. A pulse blanker may be implemented in a simple analog way (suboptimal) or in a more advanced digital approach (optimal). Analog pulse blanking is done based on power measurements, careful AGC (fast enough) operation, and proper threshold setting of the ADC. Digital implementation requires multiple bit ADC, processing of ADC output levels in real time, and optimal parameter estimation.

The performance in terms of C/N_0 of blanking pulse-type interference is given by the expression [13.59]

$$(C/N_0)_{\text{eff}} = 39.5 + 20 \log(1 - \text{PDC}_B) - 10 \log \left(1 - \text{PDC}_B + \sum_{i=1}^N 10^{\frac{R_i}{10}} \right) \quad (13.39)$$

$$R_i = P_i + 97 \text{ dBm} + 10 \log(dc_i),$$

where PDC_B is the pulse duty cycle of blanker for strong pulses, N is the number of low-level undesired pulses, P_i is the peak-received power of low-level pulses, and dc_i is the duty cycle of low-level pulses.

13.3.2 Galileo Receivers

The architecture of Galileo receivers is basically not too different from GPS receivers. Depending on the target Galileo service (open service, commercial service, public regulated service), the receiver is potentially a single- or a multifrequency wide-band receiver. Aside from the PRS service, it seems to be clear that a Galileo receiver will always make use of a GPS single- or multifrequency implementation. Thus, it will be in most cases not a standalone Galileo but a hybrid GPS/Galileo receiver.

Thus, in implementing the Galileo functions into the RF front end and in the SPU, a basic decision has to be made by the designer about the Galileo services to be used and the frequency bands to be processed. After this decision is taken, the Galileo specific signal processing and data modulation have to be implemented. The RF front-end chain has to be adapted to the bandwidth of the specific Galileo signals. In particular, the bandwidth requirements are very high, if we consider, for example, the AltBOC(15,10) as a unified signal on E5 or the BOC_{cos}(15,2.5) which is the PRS component of the E1. The AltBOC(15,10) has a bandwidth of 50 MHz (the entire signal components are transmitted in a 90 MHz bandwidth) and if this is Nyquist sampled, it will result in an extremely fast digital data stream of > 100 Mbit/s (baseband sampling) to be digitized and processed in the SPU. The design of the DLLs and PLLs will be more or less conventional.

A Galileo high-end receiver will potentially make use of several front ends where the RF bandwidth is dependent on the selected signal (E1, E5a+b, E6) combination. As outlined earlier, RF semiconductor technologies are available to build a highly integrated RF unit. New challenges arise mainly in the digital part of the receiver (including the ADC) because of high digital data rates to be processed. Special requirements in the digital domain are the acquisition and processing of various BOC(n,m)-type signals. Open service Galileo receivers will be focused to acquire and process the composite MBOC = $(\frac{10}{11})$ BOC(1,1) + $(\frac{1}{11})$ BOC(6,1). Additionally, several data-free carriers (pilot signals) have to be processed. Robust detector designs are necessary to avoid ambiguities which could result from side-lobe tracking of the binary offset carrier (BOC) autocorrelation functions. A basic method is called *bump-jumping* which works with five correlation points. Another new element is the

decoding of the various data messages in the Galileo data structure. After frame synchronization on some of the new message types (e.g., F/NAV), de-interleaving, Viterbi decoding, data decryption, and CRC check-sum computation have to be performed. Another issue is to perform hybrid GPS/Galileo positioning making use of the Galileo/GPS time offset (GGTO).

In the Galileo, the coexistence of the E5 and E6 signals with other service signals has to be considered. As in the case of GPS, L5-pulsed interference from DME/TACAN and/or JTIDS/MIDS is to be expected. In some European countries, high-power civil and military pulsing radars are transmitting into the E6 band. Additionally, amateur TV transmissions are locally present. In order to utilize E5 and E6, the techniques of analog or digital pulse blanking and other interference mitigation methodologies have to be implemented.

In a multifrequency GPS/Galileo receiver, ephemeris, positioning, and timing data will be available from many different satellite sources. Simplified acquisition strategies could thus be implemented. If the receiver position, ephemeris data, and clock error are known from, for example, GPS, then Galileo signal search could be skipped by the calculation of the Doppler and code phase from an available GPS navigation solution. If this is not possible, handover to Galileo signal acquisition unit has to be done.

13.3.3 GLONASS Receivers

The historical difference between GPS and GLONASS was that GLONASS used a frequency division multiple access (FDMA) scheme in contrary to GPS which uses code division multiple access (CDMA). In the context of GLONASS modernization, it is planned to transmit various CDMA signals (like GLONASS-K1, K2, KM) step by step. The implementation will need several years until 2025+. However, because of obvious backward compatibility issues, the FDMA signals will also be present in future.

The processing of FDMA signals in a receiver on the analog and digital levels has a basic impact on the receiver design and performance [13.60]. In the case of FDMA, the RF front end has a much higher complexity than that of CDMA. The number of components is higher and the frequency synthesizer is much more difficult to implement. The cost and effort of designing the GLONASS RF chip are also higher than that of the single-carrier frequency CDMA. The front end is also the driving factor of the power consumption and the form factor of the receiver. The requirements for the FDMA synthesizer [13.60] are to elaborate an optimal frequency plan, the selection of proper ref-

erence frequency/division ratios, minimize the phase noise penalty, and minimize the number of components at the same time. The integration of GPS and/or Galileo to GLONASS in the receiver adds even more complexity to the synthesizer.

A second problem area is that the designer has to minimize the group delay dispersion in the RF front end (which is present for each individual satellite) by the appropriate design of (band pass) filters which provide small delays without temperature variation. Special calibration loops are necessary for high precision. The presence of differential group delay between GLONASS satellites will degrade the performance of the navigation solution.

13.3.4 BeiDou/Compass Receivers

Due to signal characteristics of BeiDou-1 service signals, BeiDou-1 receivers were at least of backpack size, which is not sufficient for handheld versions in comparison to the other two existing systems. However, because of clear similarities in the signal structures of BeiDou-2 and Galileo, the architectural design of a BeiDou-2 receiver became similar to a Galileo receiver. In June 2004, the first generation of handheld BeiDou-2 receivers came out of Taipei by miniaturization of devices.

Similar to GLONASS receivers, the use of a specific reference frame in BeiDou which is different from GPS/Galileo may be a problem in designing a multisystem receiver. The regional short message service of BeiDou, which allows the user and the station to exchange short message (currently 120 Chinese characters per message), adds complexity to the receiver, and therefore potentially higher costs. Furthermore, a vague policy of BeiDou signal design, for example, modulation schemes, and navigation message contents, and formats was also a barrier in GNSS receiver industry to get into the market. However, the BeiDou B1 signal interface control document, released in December 2012, has boosted new development activities of worldwide GNSS chipset and receiver manufacturers [13.61]. An initial and logical step in developing such a device or component is to reuse existing technology widely proven by previous design examples. For an example, STMicroelectronics introduced the so-called Teseo-II chip which was designed based on STA8088 originally built for GPS, GLONASS, and Galileo. The final solution employs two chips (the Teseo-II itself and an additional STA5630 tuner) and can do positioning with BeiDou only or BeiDou + GLONASS. The receiver software integrates the new constellation with the existing ones (GPS, Galileo, GLONASS, and regional systems), such that the selection of various

GNSS configurations is actually possible [13.62]. For another example, Broadcom offers the GNSS location chip with BeiDou support. They have introduced the BCM47531, a GNSS chip that generates positioning data from five satellite constellations simultaneously – GPS, GLONASS, QZSS, SBAS, and BeiDou. The newly added BeiDou constellation increases the satellite visibility to a smartphone, enhancing navigation accuracy, particularly in urban settings where buildings and obstructions can affect performance.

It is noted that transportation authorities of China built up a mandate that certain commercial vehicles in parts of the nation should use the BeiDou system. This is similar as the activity of Russian government for GLONASS. This and the completion of BeiDou constellation will be an important driver to increasing of a BeiDou receiver-equipped navigation system on their market [13.63].

13.3.5 Military GPS Receivers

After the initial operational capability (IOC) of GPS was reached in December, 1993, military receivers which were delivered to US and NATO forces were furnished with the precise positioning service-security module (PPS-SM). PPS-SM was a first generation crypto module [13.64] to allow a military receiver to get access to the Y-code. The security module was a central processing chip in the receiver. It provided, on the one hand, the interface to the crypto key loader at the receiver housing. On the other hand, it steered the distribution of cryptographic information inside the receiver. Military receivers of this generation made use of P-code reference generators in all of the channels. In order to add a Y-code capability to each of the channels, the crypto information was directed to the P-code channels via so-called auxiliary output chips (AOCs) [13.65]. Additionally, Selective Availability (SA) parameters were handed over to the navigation processor which allowed a receiver for authorized users (PPS) to remove SA from the navigation solution. Many military receivers were manufactured in accordance with the PPS-SM standard in the 1990s (e.g., Rockwell Collins MAGR, PLGR, GEM I-IV, Trimble FORCE, etc.).

The problem faced over the years was that all these receivers (delivered by different manufactures) came up with different interfaces, form factors and test standards. Thus, in 1998, the Chairman of the Joint Chiefs of Staff issued a deployment mandate for modular standard and new security module architecture. This new modular standard [13.64] is called the GPS receiver application module (GRAM). It includes a new security architecture which is called selective availability anti-spoofing module (SAASM).

In general, GRAM is a new DoD/industry standard for embedded GPS which is aiming at a modular and open system architecture with standardized electrical, functional, and software interfaces including standardized procedures for verification and testing. The goal is an exchangeable receiver for system integrators, that is, the same receiver architecture with different form factors. Military receivers in *six* form factors are supported: a standard electronic module, format E (SEM-E) for aviation, a versa module euro (VME) card bus implementation for Navy ships, a Personal Computer Memory Card International Association (PCMCIA) for mobile terrestrial applications, a synchronous serial in-

terface (SSI) capable form factor for handheld receivers like the defense advanced GPS receiver (DAGR) and two form factors for guided munitions.

The GRAM concept should allow for fast and cost-effective upgrades and should lead to a defined market for military GPS products. Following the public domain information given in [13.65, 66], the SAASM chip is a systems on the chip (SoC) layout where several functional silicon dies are integrated: a key data processor (second generation crypto module), an advanced MP, RAM and ROM storage units, and the code block including the digital logics for acquisition and correlation.

13.4 Technology Trends

Independent of the specific receiver (civil low-end, civil high-end, military, etc.), a general trend in GNSS receiver technology will be ongoing miniaturization. Higher VLSI will result in lower weight, smaller size, and lower power consumption. In the past, higher integration led to lower production and acquisition cost for the chips and finally the receivers. However, leading-edge VLSI (65 nm and below) is depending on market size for the class of receivers. We will see in all receiver domains hybrid receivers which will use GPS and one or more other GNSS system. Another general trend is that more correlators (up to massively parallel) will be used per channel in order to enhance sensitivity and help to mitigate multipath effects by more advanced mitigation techniques. Because many new signals are potentially available, more flexible or generic implementation concepts could become an issue.

13.4.1 Civil Low-End Trends

The civil low-end receiver technologies are deployed into the mass market where 1 billion (2013 estimate) GPS chips are in use in mobile phones, automotive equipment, mobile computers, and other consumer products. Implementation of GNSS ASICs in 65 nm CMOS has happened. The question is if and when will the step to next technology node happen, for example, 45 nm CMOS. In this receiver categories, power efficiency and area efficiency are main drivers of the implementation. Integrated single chips where analog RF and digital CMOS are integrated together are already a standard. In order to enhance the sensitivity down to signal levels of -190 dBW, massively parallel correlators or comparable FFT techniques are in use for acquisition.

Apparently, a single-frequency dogma is persistent: currently only single-frequency L1 chips and/or dies are

developed. The main argument against dual frequency is that the analog CMOS RF modification necessary for, for example, L2 or L5 is a significant cost driver (cost factor of 2). For the digital subsystem, no major cost impact exists. Overall, a more expensive chip will result which is not demanded by the mass market customers. It is argued that no requirement for higher performance by adding a second frequency is present in the mass market.

Another problem area is the question how many GNSS systems should be supported on the same chip. This question cannot be finally answered. It has to do with the computational limits and finally power consumption of the single chip. A trend seems to be that, in any way, there will be a GPS L1 implementation as a basis for the chip. A regional treatment of customer expectations could lead to the design option that we will have GPS/Galileo for the European market, GPS/GLONASS for the Russian, and GPS/BeiDou integrated hybrid chips for the Chinese markets.

Another open question at the low end is: If really an upcoming competition by pure software solutions against hardware implementation would take place?

13.4.2 Civil High-End Trends

At the civil high-end, highest power efficiency and high area efficiency are not the main design drivers. Although the large-scale integration proceeds also in this area, the degree of integration is lagging behind the low end by one or two innovation cycles. The trend is clearly to use all GNSS systems on all available frequency bands, that is, to provide a multisystem and multifrequency receiver. Such receivers will utilize more than 220 channels. They have a complex front-end ASIC, a powerful digital ASIC and an MP to run differ-

ent application specific software for real-time kinematic (RTK) positioning or PPP (precise point positioning). Most receivers are optimized for high-precision carrier-phase measurements with mm precision. A more precise oscillator is a requirement to keep the phase-noise effects small in working with small tracking loop bandwidth. In order to enter new markets, these high-end receivers will be provided in the form of modules. At the high end, the software-defined radio based on a general purpose computer has potential market opportunities in cases where high adaptability and flexibility with respect to future ICD updates and changes of the signal formats are the requirements.

13.4.3 Trends in Military and/or Governmental Receivers

The current generations of GRAM/SAASM receivers which are mandatory since 2002 in the USA and their allied forces make all use of the P(Y)-code. Since the launch of the first GPS IIR-M, the new military M-code is transmitted. Thus, it is quite straightforward for the DoD to develop receivers which have M-code capa-

bility. The development program is called modernized user equipment (MUE). A request for proposal (RFP) was launched in 2004 by the US government with the goal to develop a so-called YMCA engine. The RFP called for a proof of concept and the development of ASICs and/or modules plus cost estimation. Since October 2007, three contracts had been awarded to develop the YMCA [13.67] receiver: ground-based-GRAM-M SSI (Rockwell Collins), GRAM SEM-E/M Card (Raytheon), single YMCA ASIC/SoC (L-3/interstate electronics). Besides this GRAM form-factor, it was announced to develop a common GPS module (CGM) which is making use of the latest semiconductor technology. Another interesting trend is the development of the MicroDAGR by Rockwell Collins. The MicroDAGR, with 0.175 kg, is a smaller version of the handheld DAGR which has a weight of 0.450 kg. It is a fusion of a military handheld GPS with mobile multimedia features. Some of these features are a digital camera, MP3 player, and a colored touch screen which are integrated. The motivation for the MicroDAGR is to prevent the use of small-sized commercial GPS receivers by troops in military theatres.

13.5 Receiver Types

In this section, we give a brief and systematic overview about the main types of GNSS receivers. The section is written without making any claim for completeness because over the years so many different specialized receivers were developed.

13.5.1 Navigation Receivers Handheld

Navigation receivers of handheld type have the appearance of a mobile radio. They make use of an integrated antenna, an integrated battery, a simple keyboard, and a monochrome or colored LCD display. On the architecture side in the civil domain, 12-channel L1 C/A-code receivers are in use. In the future, other GNSS systems will be integrated on the L1 frequency on a regional market-oriented basis. Most of them have software for waypoint navigation and some have a real-time DGPS capability. Governmental-military receivers make use of a security module like PP-SM or SAASM. Handheld receivers are characterized by low cost (about 200 US \$ at the low end), and are used in the land, marine, and aviation domain.

13.5.2 Navigation Receivers Non-Handheld

The appearance of nonhand-held navigation receivers is, in most cases, a sensor (black box) with specific

interfaces for marine, land, and aviation applications. In the civil market, 12-channel L1 C/A-code GPS receivers are available. For these receivers, small form factor and low power consumption are usually not a design driver, because they are used in large platforms like ships or commercial aircraft. In the military domain, dual-frequency P(Y)-code receivers are standard until now. Especially, in commercial aviation, these receivers are used in safety-critical operations like area navigation and precision landing. For this purpose, receivers have to undergo specific qualification and certification procedures such as MIL-STD 810, Federal Aviation Administration (FAA) TSO C-129, ARINC 743 A, RTCA DO-217, and so on. Most navigation receivers have a requirement for autonomous integrity monitoring schemes like receiver autonomous integrity monitoring (RAIM). Depending on the user community, these receivers are compatible with a variety of interfaces like Mil-Bus 1553, ARINC-429, National Marine Electronics Association (NMEA) 0183, Radio Technical Commission for Maritime Services (RTCM) SC 104, and others. Usually, a specific user dynamics in terms of linear velocity, acceleration, and jerk for the signal dynamics is specified (500–1000 m/s, 4–9 g, 4–10 g/s). Military navigation receivers fulfill a specified jamming resistance in terms of the jamming-to-signal ratio (J/S). Because of these ad-

vanced certification and standardization requirements, the acquisition cost of nonhand-held receivers is on the 10 000–30 000 US \$ level.

13.5.3 Engines, OEM Modules, Chips, and Dies

GNSS engines, original equipment manufacturer (OEM) modules, chips, and dies are semifinished products. No housing, power supply, and control and display unit are attached to these receivers. This class of GNSS products is integrated into higher level navigation systems which are sold in a specific market, for example, car navigation systems. Most of these receivers are parallel channel C/A-code L1 units until now. Multifrequency and multisystem modules are also available in the current years. These semifinished products usually provide flexible hardware and software interfaces which allow to output all the measured raw data and the entire navigation message.

13.5.4 Time Transfer Receivers

Timing receivers usually make use of a low-end 12-channel GPS L1, C/A-code chip. The timing system itself is based on a more precise clock/oscillator like, for example, a rubidium. Specific optimal estimation algorithms are applied in order to determine the clock offset and frequency offset from GPS time and synchronize to Coordinated Universal Time (UTC) (United States Naval Observatory (USNO)). Early time transfer receivers provided only a single channel because only one unknown (the time offset) has to be estimated. The user world geodetic system (WGS) 84 coordinates of the antenna had to be input. Because the GNSS chips developed in recent years make use of multiple channels, time transfer receivers with a single channel are not provided anymore. Thus, timing receivers have also a positioning capability nowadays.

13.5.5 Geodetic Receivers

Modern geodetic or surveying GNSS receivers have a multisystem, multifrequency design. They are able to track all current and future GNSS signals and thus provide (based on several high-end GNSS modules) more than 400 channels. Currently, most of these receivers have a semicodeless capability to make available at least P(Y)-code pseudoranges and carrier phases on L2. Surveying receivers are optimized to provide high-precision carrier-phase measurements with millimeter accuracy. This high accuracy implies the use of a more advanced antenna with a stable phase-center and a means to suppress the ground multipath. It is

necessary that the surveying devices are adapted to the typical field environment of surveyors. From the form-factor side, the GNSS modules are typically integrated together with different radio modems (for the reception of RTK correction data) in the antenna pot. The antenna pot is mounted on top of a vertical rod. The rod includes a leveling device which is used to center the antenna exactly above a survey mark on ground. Attached to the rod is a CDU for system control functions. The receiver usually includes a large memory for storing all the measured data for post-processing purpose. Present surveying receivers have a RTK capability on the software level. This allows the receiver to process various differential correction formats. Sometimes the receiver is integrated together with other surveying instruments like theodolites and laser ranging equipment to a total station. The acquisition cost of GNSS surveying systems is on the level of multiple 10 000 US \$.

13.5.6 Space Receivers

Since a decade, the use of GNSS receivers on-board of Earth satellites became a standard. GNSS on spacecraft is an inexpensive way to perform autonomous orbit and time determination (Chap. 32). Apart from the higher Doppler shifts of ± 40 kHz during orbital motion at about 8000 m/s, the GNSS positioning scenario on an LEO platform is not so different from an Earth-fixed user. This situation changes, if the receiver is flown on high-flying satellites in geostationary Earth (GEO) or highly elliptical (HEO) orbits. Because the user in this type of orbit is above the GNSS constellation, it can only receive satellites which transmit on side-lobes from the other side of the Earth. A spaceborne receiver has some special features; because of the high Doppler shift, specific acquisition algorithms are necessary. For autonomous orbit determination, nonlinear Kalman filters are used. Because the user platform is flying on a high altitude, GNSS satellites are visible under the negative mask angle below the horizon. For specific missions, the Earth itself as a shading body has to be considered in the visibility analysis. Because many platforms are not flown in an Earth-pointing attitude, the receiver needs the capability (or option) to work with two GNSS antennas in order to get independent of the spacecraft attitude. Other features could be the use of radiation-hardened components and the provision of interfaces to specific spacecraft data bus systems.

13.5.7 Attitude Determination Receivers

For the attitude determination of a user platform, multiple antenna systems are used. The key issue of attitude determination is to measure precise carrier phases be-

tween the antenna phase centers while knowing the baseline (the distance) between them (Chap. 27). This opens the possibility of determining pitch, roll, and yaw of the platform with an accuracy of $0.1\text{--}0.01^\circ$ depending on the base-line length, the carrier-phase multipath, and noise errors. In many GPS attitude determination systems, four antennas are used. Behind each antenna, a receiver is located with say a bank of six channels each. The different receivers are synchronized between

each other by using a common crystal oscillator. Attitude receivers are used on terrestrial and spaceborne platforms.

Acknowledgments. The authors would like to thank to Dr. Xingqun Zhan, Professor and Associate Dean of the School of Aeronautics and Astronautics at Shanghai Jiao Tong University, Shanghai, China for his support on early BeiDou receiver development in Sect. 13.1.5.

References

- 13.1 B.W. Parkinson: Introduction and heritage of NAVSTAR the global positioning system. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington 1996) pp. 3–28
- 13.2 P.C. Ould, R.J. van Wechel: All-digital GPS receiver mechanization, *Navigation* **28**(3), 178–188 (1981)
- 13.3 Data Sheet on MX 4200 12 channel up-grade (Magnavox Electronic System Company, Torrance 1994)
- 13.4 A.J. van Dierendonck, P. Fenton, T. Ford: Theory and performance of narrow correlator spacing in a GPS receiver, *J. Inst. Navig.* **39**(3), 265–284 (1992)
- 13.5 S. Mikkola: Generalized Development Model (GDM) (Institute of Navigation Navigation Museum, ION, Virginia) http://www.ion.org/museum/item_view.cfm?cid=7&scid=9&iid=9
- 13.6 Receiver 3 A Configuration, Product Information Sheet (Rockwell International, Collins Government Avionics Division, Cedar Rapids 1987)
- 13.7 R. Hoeh, R. Bartholomew, V. Moen, K. Grigg: Design, capabilities and performance of a miniaturized airborne GPS receiver for space applications, *Proc. IEEE PLANS*, Las Vegas (1994) pp. 1–7
- 13.8 TI-4100 Owner's Manual (Texas Instruments, Lewisville 1983)
- 13.9 A.J. van Dierendonck: GPS receivers. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington 1996) pp. 329–407
- 13.10 Global Positioning Product Handbook (GEC-Plessey Semiconductors, Plymouth 1996)
- 13.11 J. Ashjaee, R. Lorenz: Precision GPS surveying after Y-Code, *Proc. ION GPS*, Albuquerque (1992) pp. 657–659
- 13.12 A.J. van Dierendonck: Understanding GPS receiver terminology: A tutorial on what those words mean, *Proc. Int. Symp. Kinemat. Syst. Geod. Geomat. Navig.*, KIS94, Banff (1994)
- 13.13 C.C. Counselman: Method and system for determining position using signals from satellites, US Patent 4 667 203A (1982), Aero Service Div.
- 13.14 K.T. Woo: Optimum semi-codeless carrier phase tracking of L2, *Navigation* **47**(2), 82–99 (2000)
- 13.15 J.M. Fraile-Ordóñez, G.W. Hein, H. Landau, B. Eissfeller, A. Jansche, N. Balteas: First experience with differential GLONASS/GPS positioning, *Proc. ION GPS*, Albuquerque (1992) pp. 153–158
- 13.16 P. Misra, P. Enge: *Global Positioning System – Signals, Measurements, and Performance*, 2nd edn. (Ganga-Jamuna, Lincoln 2006)
- 13.17 Inside GNSS: NovAtel confirmed for long-term Galileo contract (2007) <http://www.insidegnss.com/node/247>
- 13.18 IFEN GmbH: Multi-GNSS navigation test receiver, NAVX-NTR, Data-Sheet, 2012
- 13.19 A. Simsky, J.M. Sleewaegen, W. de Wild, F. Wilms: Galileo receiver development at septentrio, *Proc. ENC GNSS*, Munich (2005) pp. 1–14
- 13.20 A. Ruegamer, I. Suberviola, F. Foerster, G. Rohmer, A. Konovaltsev, N. Basta, M. Meurer, J. Wendel, M. Kaindl, S. Baumann: A Bavarian initiative towards a robust Galileo PRS receiver, *Proc. ION GNSS*, Portland (2011) pp. 3668–3678
- 13.21 Unicore Communications, Inc.: <http://www.unicorecomm.com/>
- 13.22 OLinkStar Co., Ltd.: <http://www.olinkstar.com/>
- 13.23 ComNav Technology Ltd.: <http://www.comnavtech.com/>
- 13.24 NovAtel: GPSAntenna Model 501, User Manual (NovAtel) <http://www.novatel.com/assets/Documents/Manuals/om-20000001.pdf>
- 13.25 J.D. Kraus, K.R. Carver: *Electromagnetics* (McGraw-Hill, Tokyo 1973)
- 13.26 EC: Galileo Overall Architecture Definition – GALA (European Commission, Brussels 2000) Gala-Aspid087
- 13.27 E. Levine: Overview of GPS antennas, *Proc. COMCAS*, Tel Aviv (2009) pp. 1–4
- 13.28 D. Reynolds, A. Brown, A. Reynolds: Miniaturized GPS antenna array technology and predicted anti-jam performance, *Proc. ION GPS*, Nashville (1999) pp. 777–786
- 13.29 W. Kunysz: A three dimensional choke ring ground plane antenna, *Proc. ION GPS/GNSS*, Portland (2003) pp. 1883–1888
- 13.30 EADS Astrium, GNSS Evolution Programme: Assessment of the Use of C-Band for GNSS, Final Report (GNSS-CBA-TN-ASTD-00023, June 2009)
- 13.31 Data Manual ADC12D1800RF 12 Bit, Single 3.6 GSPS RF Sampling ADC (Texas Instruments, 2015)
- 13.32 H.T. Friis: Noise figures in radio receivers, *Proc. IRE* **32**(7), 419–422 (1944)
- 13.33 T. Felhauer: On the impact of RF front-end group delay variations on GLONASS pseudorange accu-

- racy, Proc. ION GPS, Kansas City (ION, Virginia 1997) pp. 1527–1532
- 13.34 J.B. Neumann, M. Bates, R.S. Harvey: GLONASS receiver inter-frequency biases – Calibration methods and feasibility, Proc. ION GPS, Nashville (ION, Virginia 1999) pp. 329–338
 - 13.35 T. Pany: *Navigation Signal Processing for GNSS Software Receivers* (Artech House, Norwood 2010)
 - 13.36 T. Pany, B. Eissfeller: Code and phase tracking of generic PRN signals with sub-nyquist sample rates, *Navigation* **51**(2), 143–159 (2004)
 - 13.37 R.G. Vaughan, N.L. Scott, D.R. White: The theory of bandpass sampling, *IEEE Trans. Signal Proc.* **39**(9), 1973–1983 (1991)
 - 13.38 R.G. Vaughan, N.L. Scott, D.R. White: Analog-to-digital converters and their applications in radio receivers, *IEEE Commun. Mag.* **33**(5), 39–45 (1995)
 - 13.39 D.M. Akos: A Software Radio Approach to Global Navigation Satellite System Receiver Design, Ph.D. Thesis (Ohio Univ., Athens 1997)
 - 13.40 J.H. Won: Studies on the Software-Based GPS Receiver and Navigation Algorithms, Ph.D. Thesis (Ajou Uni., Suwon 2004)
 - 13.41 A.R. Pratt, J.A. Avila-Rodriguez: Time and amplitude quantization losses in GNSS receivers, Proc. ION GPS, Savannah (ION, Virginia 2009) pp. 3179–3197
 - 13.42 C.J. Hegarty: Analytic model for GNSS receiver implementation losses, Proc. ION GPS, Savannah (2009) pp. 3165–3178
 - 13.43 D. Borio: A Statistical Theory for GNSS Signal Acquisition, Ph.D. Thesis (Politecnico di Torino, Turin 2008)
 - 13.44 J.J. Spilker Jr., F.D. Natali: Interference effects and mitigation techniques. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington 1996) pp. 717–771
 - 13.45 Chip Scale Atomic Clock, SA.45s CSAC, Data Sheet (Symmetricom, San Jose 2011)
 - 13.46 D. Allan: Statistics of atomic frequency standards, *Proc. IEEE* **54**(2), 221–230 (1989)
 - 13.47 A.R. Pratt: G-effects on oscillator performance in GPS receivers, *Navigation* **36**(1), 63–75 (1989)
 - 13.48 F.L. Walls, J.O. Gary, A. Gallagher, L. Sweet, R. Sweet: Time domain frequency stability calculated from the frequency domain: An update, Proc. 4th Eur. Freq. Time Forum, Neuchatel (1990) pp. 197–204
 - 13.49 M. Irsigler, B. Eissfeller: PLL tracking performance in the presence of oscillator phase noise, *GPS Solutions* **5**(4), 45–57 (2002)
 - 13.50 P.S. Otellini: (Investor Meeting 2012, Intel, Santa Clara, 2012)
 - 13.51 G.E. Moore: No exponential is forever: But *forever* can be delayed!, Proc. IEEE Int. Solid-State Circuits Conf., San Francisco (2003) pp. 20–23
 - 13.52 H. Kurz: Visions of semiconductors, Proc. Munich Satell. Navig. Summit, Munich (2010)
 - 13.53 F. Schwiertz, J.J. Liou: *Modern Microwave Transistors – Theory, Design and Performance* (John-Wiley, Hoboken 2003)
 - 13.54 G. Kappen, T.G. Noll: Application specific instruction processor based implementation of a GNSS receiver on an FPGA, Proc. DATE, Munich (2006) pp. 1–8
 - 13.55 TMS 320 DSP Product Overview (Texas Instruments, 1998)
 - 13.56 S. Furber: *ARM System-on-Chip Architecture* (Addison Wesley, New York 2000)
 - 13.57 J. Ashjaee: GPS: The challenge of a single chip, *GPS World* **12**(5), 24–27 (2001)
 - 13.58 M. Tran, C. Hegarty: Receiver algorithms for the new civil GPS signals, Proc. ION NTM, San Diego (ION, Virginia 2002) pp. 778–789
 - 13.59 C. Hegarty, A.J. van Dierendonck, D. Bobyn, M. Tran, T. Kim, J. Grabowski: Suppression of pulsed interference through blanking, Proc. IAIN World Congr. ION AM, San Diego (ION, Virginia 2000) pp. 399–408
 - 13.60 O. Leisten, R. Hasler, M. Malsor: Design and performance of a miniature GPS/GLONASS receiver, *Micro. Eng. Eur.*, 33–38 (Dec./Jan. 1992)
 - 13.61 BeiDou Navigation Satellite System Signal In Space Interface Control Document (Test Version) (China Satellite Navigation Office, Beijing 2012)
 - 13.62 F. Pisoni, P.G. Mattos: A BeiDou hardware receiver based on the STA8088 chipset, Proc. ICL-GNSS 2013, Turin (2013) pp. 1–6
 - 13.63 InsideGNSS: China Mandates Use of BeiDou GNSS on Some Commercial Vehicles, InsideGNSS (2013) <http://www.insidegnss.com/note/3356>
 - 13.64 E. Emile, S.L. Saks: GPS receiver application module (GRAM) open system architecture (OSA) for next-generation DoD GPS receivers, Proc. ION NTM, Long Beach (1998) pp. 297–307
 - 13.65 K. Goussak, T. Kusserow, B. Goblish: Review and analysis of the selective availability anti-spoofing module (SAASM) card integration program (SCIP), Proc. ION AM, Denver (ION, Virginia 1998) pp. 585–592
 - 13.66 H. Fruehauf, S. Callaghan: SAASM and direct P(Y) signal acquisition, *GPS World* **13**(7), 24–33 (2002)
 - 13.67 R. DiEsposti: Proposed operations concepts and flexible UE architectures for modernized user equipment SIS utilization for transition from test mode to IOC through FOC, Proc. ION NTM, San Diego (ION, Virginia 2007) pp. 548–560

Signal Processing

14. Signal Processing

Jong-Hoon Won, Thomas Pany

In this chapter digital signal processing of a global navigation satellite system (GNSS) receiver is presented. It provides a high-level block diagram as well as detailed descriptions for all the internal functions of a modern digital GNSS receiver, focusing on signal acquisition and tracking, time synchronization, navigation data bit demodulation and decoding, and measurement generation. Also, several issues on the processing of upcoming GNSS signals, which may have new features like a binary offset carrier (BOC) modulation, data/pilot channels, primary/secondary codes, and so on are addressed. Furthermore, advanced topics in designing modern digital GNSS receivers such as tracking of the global positioning system (GPS) P(Y)-code, various combined processing schemes, Kalman filter-based signal tracking loops, and a vector-tracking approach are also presented.

14.1	Overview and Scope	402	14.4	Signal Tracking	413
14.2	Received Signal Model	403	14.4.1	Architecture.....	413
14.2.1	Generic GNSS Signal.....	403	14.4.2	Tracking Loop Model.....	414
14.2.2	Signal Model at RF and IF.....	404	14.4.3	Correlators.....	415
14.2.3	Correlator Model.....	404	14.4.4	Discriminators.....	415
14.3	Signal Search and Acquisition	406	14.4.5	Loop Filters.....	417
14.3.1	Test Statistics.....	406	14.4.6	NCO and Code/Carrier Generator.....	419
14.3.2	Acquisition Module Architecture.....	408	14.4.7	Aiding.....	421
14.3.3	Coherent Integration Methods.....	409	14.4.8	Switching Rule.....	421
14.3.4	Search Space.....	410	14.4.9	BOC Tracking.....	422
14.3.5	Acquisition Performance.....	411	14.4.10	Tracking Performance.....	422
14.3.6	Handling Data Bits and Secondary Codes.....	412	14.5	Time Synchronization and Data Demodulation	424
			14.5.1	Bit/Symbol Synchronization.....	425
			14.5.2	Data Bit/Symbol Demodulation.....	426
			14.5.3	Frame Synchronization.....	427
			14.5.4	Bit Error Correction.....	427
			14.5.5	Data Extraction.....	428
			14.6	GNSS Measurements	428
			14.6.1	Code Pseudorange.....	428
			14.6.2	Carrier Phase.....	431
			14.6.3	Doppler.....	433
			14.6.4	Signal Power.....	434
			14.7	Advanced Topics	434
			14.7.1	Tracking of GPS P(Y).....	434
			14.7.2	Generic Data/Pilot Multiplexing Approach.....	435
			14.7.3	Combined Processing of Data and Pilot Signals.....	436
			14.7.4	Combined Processing of Code and Carrier.....	436
			14.7.5	Carrier Tracking Kalman Filter.....	437
			14.7.6	Vector Tracking.....	438
			References		440

14.1 Overview and Scope

The signal processing unit of a GNSS receiver processes the combined signal from all satellites plus noise and interference. This is done on one or more frequency bands. Typically, the receiver has dedicated units to process the signal from the individual satellite and frequency bands – those units are called channels.

The received signal power from the satellite is very low, usually -130 dBm or less, and the amplitude of the received signal is much smaller than the one of the received noise. One speaks of the fact that the signal is buried in the noise. Furthermore, the satellites transmit on the same carrier frequency and signals from different satellites overlap. As a matter of fact the typical received signal after amplification, downconversion and analog-to-digital conversion (ADC) looks like the data shown in Fig. 14.1. The GNSS signals are not directly recognizable within this data stream, and the signal samples look purely random.

To detect and track the GNSS signals, the receiver employs the *auto-correlation* principle. It generates a transmitted GNSS signal copy of a single satellite inside the receiver and correlates this replica signal with

the received signal. If the signal parameters in terms of code phase ($=$ signal delay τ) and Doppler shift f_d match reasonably well, the correlation value increases as shown in Fig. 14.2. The correlation is realized as an *integration* of the product of received and replica signal. Thus the terms correlation and integration are used synonymously within this chapter.

The correlation brings the signal for the considered satellite above the noise floor, separates this signal from the other satellite signals and provides an estimate of the signal parameters including τ and f_d .

Figure 14.3 illustrates the schematic block diagram of a generic GNSS receiver focusing on the internal functions of a signal processing channel. GNSS signals received at the antenna are filtered, amplified, downconverted to an intermediate frequency (IF) by the radio frequency (RF) signal chain in the front end, and then applied to an ADC at the end of the RF front end to obtain digital samples. The *digital processing* of a GNSS signal in a channel starts with the detection that a signal is present (acquisition). During acquisition coarse estimates of the code delay and Doppler of the signal are determined in feed-forward manner, and then the channel switches to code and carrier tracking to refine the estimates in a feedback structure.

For tracking, several measures are necessary to ensure accuracy and stability of the tracking process. Tracking is based on generating an internal replica of the received signal and the replica is generated itself by several code and carrier numerically-controlled oscillators (NCOs). The NCO counters are converted into geometrical meaningful values (code/carrier pseudorange, Doppler) and are passed together with a signal power estimate to the navigation processor. The received signal power is permanently monitored. If it drops below a certain threshold, the channel declares a loss of lock and restarts with acquisition. During

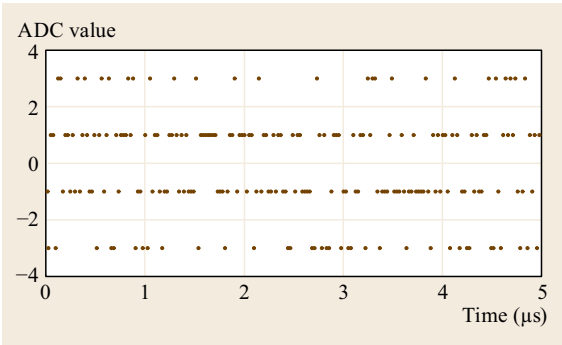


Fig. 14.1 Output of a 2 bit ADC receiving GPS C/A signals plus thermal noise

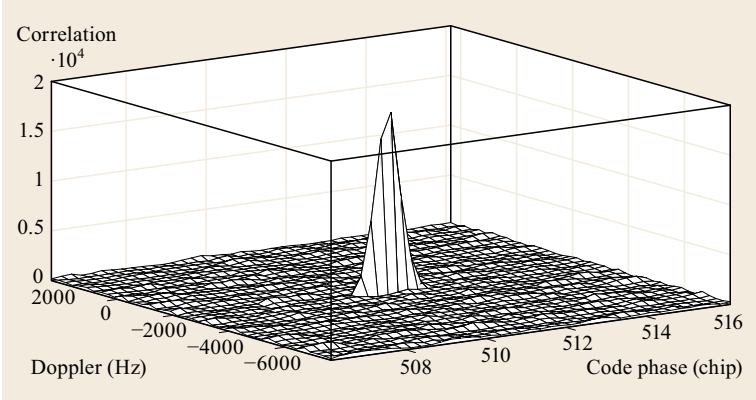


Fig. 14.2 GPS C/A correlation function computed for PRN1 during signal acquisition

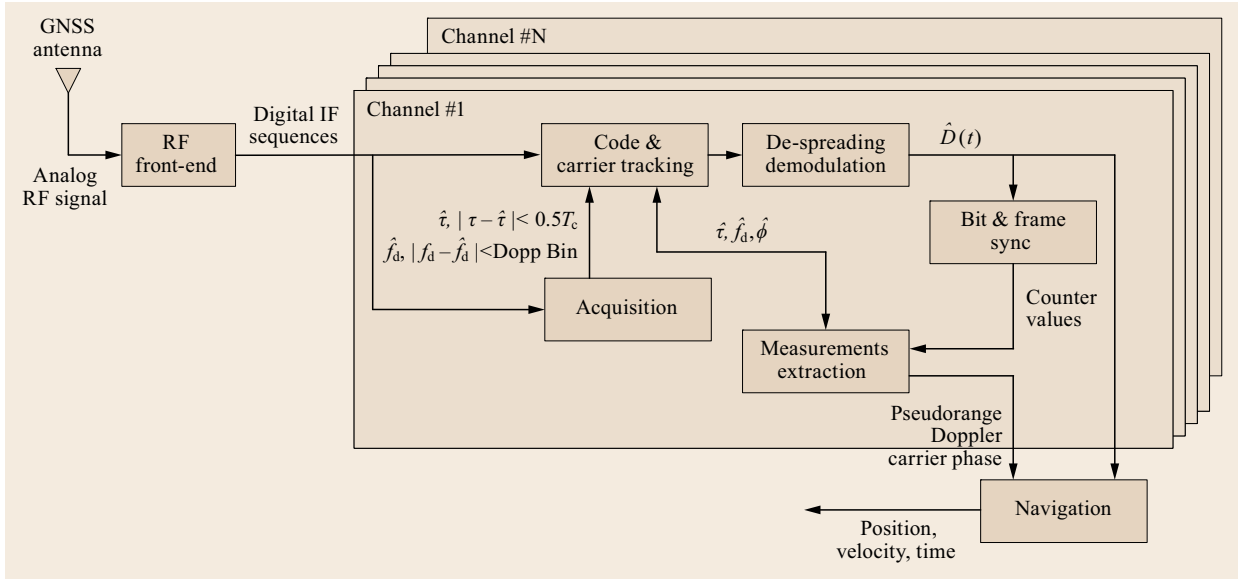


Fig. 14.3 Block diagram of internal functions in a generic GNSS receiver architecture

tracking the channel synchronizes to the broadcast navigation data message and decodes it. The decoded bit information contains the satellite's ephemeris and almanac, system time information, and meteorological parameters. The information from code and carrier

tracking blocks together with time synchronization information is used to generate the primary measurements of GNSS. Using these, finally, the navigation unit computes the GNSS navigational equation to obtain the user position, velocity and timing (PVT) information.

14.2 Received Signal Model

This section describes a model for the considered GNSS signals and of the resulting correlation values. These models form the basis for the tracking and acquisition analysis.

14.2.1 Generic GNSS Signal

Assuming a generic multiplexing scheme based on quadrature phase shift keying (QPSK) modulation, the signal transmitted at the antenna of a single GNSS satellite in one frequency band can be modeled as

$$s(t) = \sqrt{2P_c(t)}D_c(t)C_c(t)\cos(2\pi f_L t) + j\sqrt{2P_s(t)}D_s(t)C_s(t)\sin(2\pi f_L t), \quad (14.1)$$

where P is the signal power for the corresponding signal component, D is the navigation data symbol sequence (± 1 for data channels and $+1$ for pilot channels) with a symbol duration of T_{sym} in seconds, C is the spreading code sequence with a chip duration of T_c in seconds, f_L is the carrier frequency in L-band in Hz, and subscripts c and s represent the identifiers for carriers, respectively cosine and sine. This model covers the majority

of broadcast GNSS signals, and the ones not covered, such as alternative BOC (AltBOC) or code shift keying (CSK), have at least a similar structure. Also, unmodulated codes – realizing a binary phase shift keying (BPSK) modulation on a single component – or the relevant modulated codes, such as binary offset carrier (BOC) or any other code modulation for new GNSS signals, with the corresponding representative parameters can be used by modifying the spreading code sequence C .

The model given in (14.1) for QPSK signals can be easily modified for data-only signals or time-multiplexing signals. In the first case, the sine term is omitted, and for the second case, the sine term is replaced by a cosine and one has to make sure that C_c and C_s cannot be present simultaneously (i. e., $C_s(t)C_c(t) = 0$).

The received signal for N visible satellites at the end of a receiver antenna can be modeled as

$$r(t) = \sum_{i=1}^N a_i e^{j\phi_{0,i}} s_i(t - \tau_i) + n(t) \quad (14.2)$$

where s_i is the signal from the i th visible satellite, a_i is the attenuation factor for the signal power of the corresponding satellite, and $n(t)$ is the additive noise component. The signal propagation time from the i th satellite to the receiver is given by τ_i expressed in seconds and $\phi_{0,i}$ is a phase delay in radians.

14.2.2 Signal Model at RF and IF

From (14.1) and (14.2), the received signal for a single satellite at the end of a receiver antenna can be modeled as

$$\begin{aligned} r_{\text{RF}}(t; \tau, \phi_0, f_d, A_c, A_s) \\ = A_c D_c(t-\tau) C_c(t-\tau) \cos[2\pi(f_L + f_d)t + \phi_0] \\ + j A_s D_s(t-\tau) C_s(t-\tau) \sin[2\pi(f_L + f_d)t + \phi_0] \\ + n_{\text{RF}}(t), \end{aligned} \quad (14.3)$$

where τ is the code delay (s), f_d is the carrier Doppler frequency shift (Hz), ϕ_0 is the carrier-phase delay (rad), A is the signal amplitude for the corresponding signal component taking into account the signal power as well as the attenuation factor, $n_{\text{RF}}(t)$ is the band-limited additive white Gaussian noise (AWGN) having a one-sided power spectral density (PSD) of N_0 and a bandwidth determined usually by the filter chain inside the RF front end (RF-FE) of the receiver, and the subscript RF represents the identifier for the carrier.

The received signal in (14.3) has a very high carrier frequency and most digital signal processors would have difficulty in generating such a rapidly varying digital carrier wave. Therefore, we must downconvert the frequency to something more manageable. Moreover, the signal power of the received signal is very low so that it must be increased necessarily to be processed. At the same time the effective suppression of the natural noise as well as radio frequency interference should be made. This signal conditioning process is done by a series of mixing and bandpass filtering processes at the RF-FE (Sect. 14.3.2).

Without loss of generality, after the downconversion with proper band-limited filters on multiple (or single) stages and then digitization by the ADC, a signal component of the digitized IF received signal (e.g., data channel at cosine carrier) at the end of the RF-FE can be rewritten from (14.3) in the discrete time domain given by

$$\begin{aligned} r_{\text{IF}}(k; \tau, \phi, f_d, A) = \\ AD(T_s k - \tau) C(T_s k - \tau) \cos[2\pi(f_{\text{IF}} + f_d)T_s k + \phi] \\ + n_{\text{IF}}(k) \end{aligned} \quad (14.4)$$

for $k = 0, 1, 2, \dots$,

where T_s is the sampling time interval (s) such that $t = kT_s$, and n_{IF} is the corresponding noise at IF. Note that ϕ here is the carrier-phase offset at $t = T_s k$ in addition to the Doppler shift, and τ , ϕ , f_d , and A are the signal parameters of interest as functions of time t to be estimated in the receiver signal processing for navigation purposes.

In fact, the process performed in the RF-FE consists of the frequency translation, signal amplitude magnification and out-of-band limiting. Note that the only parameter changed in the downconversion process of the RF-FE is the center frequency of the carrier wave. The delay effects caused by the RF-FE process are common in all the channels and will be eliminated by the navigation process in the form of common clock bias.

14.2.3 Correlator Model

Assuming the navigation data bit does not change in the integration time interval, the digitized received signal for a signal component (e.g., cosine) of a visible GNSS satellite can be modeled as a complex signal notation based on the signal parameters of interest [14.1]. Therefore, without the use of amplitude and navigation data bit, the locally generated replica signal at the IF can be written as

$$\hat{r}_{\text{IF}}(k; \hat{\tau}, \hat{\phi}, \hat{f}_d) = 2C(T_s k - \hat{\tau}) e^{j[2\pi(f_{\text{IF}} + \hat{f}_d)T_s k + \hat{\phi}]} \quad (14.5)$$

Here, the notation for an estimator (i.e., hat) is intentionally used because the locally generated replica signal is a combination of output of estimators for the signal parameters of interest.

The correlator is a process to compute the integration of the received signal multiplied with the locally generated replica signal for a given coherent integration time. For the sake of the simplicity, let

$$\begin{aligned} r(k) &= r_{\text{IF}}(k; \tau, \phi, f_d, A), \\ \hat{r}(k) &= \hat{r}_{\text{IF}}(k; \hat{\tau}, \hat{\phi}, \hat{f}_d), \\ \Theta &= 2\pi(f_{\text{IF}} + f_d)T_s k + \phi, \\ \hat{\Theta} &= 2\pi(f_{\text{IF}} + \hat{f}_d)T_s k + \hat{\phi}, \\ \Sigma &= \sum_{k=1}^M, \end{aligned} \quad (14.6)$$

then

$$\begin{aligned} \text{corr} \left[r_{\text{IF}}(k; \tau, \phi, f_d, A), \hat{r}_{\text{IF}}(k; \hat{\tau}, \hat{\phi}, \hat{f}_d) \right] \\ = \sum r(k) \hat{r}(k), \end{aligned} \quad (14.7)$$

where $\text{corr}(x, y)$ is the correlation function of x and y , and M is the number of samples within the integration time $T = MT_s$, which is usually shorter or equal to the navigation data bit/symbol period.

We can apply the mixing process, a form of multiplication of the received signal and the locally generated replica signal as follows

$$r(k)\hat{r}(k) = 2C(T_s k - \tau)C(T_s k - \hat{\tau}) \times \cos \Theta e^{j\hat{\Theta}} + n_{\text{IF}}(k)\hat{r}(k) . \quad (14.8)$$

Applying (14.4) and (14.5) into (14.8), and then integration for a given integration time yields

$$\sum r(k)\hat{r}(k) = \sum C(T_s k - \tau)C(T_s k - \hat{\tau})2 \cos \Theta e^{j\hat{\Theta}} + 2 \sum n_{\text{IF}}(k)C(T_s k - \hat{\tau})e^{j\hat{\Theta}} . \quad (14.9)$$

The first summation in (14.9) is given by

$$\sum C(T_s k - \tau)C(T_s k - \hat{\tau}) = R(\delta\tau) \quad (14.10)$$

with $R(\delta\tau)$ being the normalized correlation function of $C(T_s k)$. In the case where a BPSK signal is considered, like the GPS C/A code signal, the correlation function takes the form

$$R(\delta\tau) = \begin{cases} 1 - \frac{|\delta\tau|}{T_c} & \text{for } |\delta\tau| \leq T_c , \\ 0 & \text{for } |\delta\tau| > T_c , \end{cases} \quad (14.11)$$

where $\delta\tau = \tau - \hat{\tau}$ is the code delay error (s) and T_c is the code chip interval (s). For BPSK signals R has a triangular shape as shown in Fig. 14.2.

The second term in (14.9) can be expanded to two parts

$$\begin{aligned} 2 \sum \cos \Theta \cos \hat{\Theta} &= \sum \cos(\Theta - \hat{\Theta}) + \sum \cos(\Theta + \hat{\Theta}) \\ &\approx \sum \cos(\Theta - \hat{\Theta}) , \end{aligned} \quad (14.12)$$

$$\begin{aligned} 2 \sum \cos \Theta \sin \hat{\Theta} &= \sum \sin(\Theta - \hat{\Theta}) - \sum \sin(\Theta + \hat{\Theta}) \\ &\approx \sum \sin(\Theta - \hat{\Theta}) , \end{aligned} \quad (14.13)$$

where

$$\begin{aligned} \Theta - \hat{\Theta} &= 2\pi\delta f_d T_s k + \delta\phi \\ \Theta + \hat{\Theta} &= 2\pi(2f_{\text{IF}} + f_d + \hat{f}_d) T_s k + \phi + \hat{\phi} \end{aligned} \quad (14.14)$$

with $\delta\phi = \phi - \hat{\phi}$ is the carrier-phase error (rad), and $\delta f_d = f_d - \hat{f}_d$ is the Doppler error (Hz).

The term $\Theta + \hat{\Theta}$ in (14.14) contains a higher-order frequency component $4\pi f_{\text{IF}} T_s k$ and the corresponding components in (14.12) and (14.13) were filtered out by the summation operation (i. e., low-pass filter).

Additionally, applying the rule of products of sine and cosine into (14.12) and (14.13) considering the summation of (14.9) yields

$$\begin{aligned} \sum \cos(2\pi\delta f_d T_s k + \delta\phi) &= \sum \left(\cos(2\pi\delta f_d T_s k) \cos \delta\phi \right. \\ &\quad \left. - \sin(2\pi\delta f_d T_s k) \sin \delta\phi \right) \\ &= \frac{\sin(2\pi\delta f_d T)}{2\pi\delta f_d T} \cos \delta\phi \end{aligned} \quad (14.15)$$

and

$$\begin{aligned} \sum \sin(2\pi\delta f_d T_s k + \delta\phi) &= \sum \left(\sin(2\pi\delta f_d T_s k) \sin \delta\phi \right. \\ &\quad \left. + \cos(2\pi\delta f_d T_s k) \cos \delta\phi \right) \\ &= \frac{\sin(2\pi\delta f_d T)}{2\pi\delta f_d T} \sin \delta\phi . \end{aligned} \quad (14.16)$$

The substitution of the results in (14.10), (14.15) and (14.16) into (14.9) yields

$$\begin{aligned} \sum r(k)\hat{r}(k) &= R(\delta\tau) \frac{\sin(2\pi\delta f_d T)}{2\pi\delta f_d T} e^{j\delta\phi} \\ &\quad + 2 \sum n_{\text{IF}} C(T_s k - \hat{\tau}) e^{j\hat{\Theta}} . \end{aligned} \quad (14.17)$$

Finally, taking into account the signal amplitude, data bit and noise effect, the correlator output in (14.7) can be written as

$$\begin{aligned} \text{corr}[r_{\text{IF}}(k), \hat{r}_{\text{IF}}(k)] &= \bar{A} D R(\delta\tau) \text{sinc}(\delta f_d T) e^{j\delta\phi} + \eta \\ &= I + jQ \end{aligned} \quad (14.18)$$

with

$$\text{sinc}(\delta f_d T) = \frac{\sin(2\pi\delta f_d T)}{2\pi\delta f_d T} , \quad (14.19)$$

$$\bar{A} = \sqrt{2TC/N_0} , \quad (14.20)$$

where \bar{A} is the amplitude of the baseband signal component assuming a normalized noise component, C/N_0 is the carrier-to-noise ratio (Hz), and I and Q represent postcorrelation values in in-phase (I) and quadrature (Q) components respectively, (this is the so

called *baseband signal* component) and can be given with
by

$$\begin{aligned} I &= \bar{A} D R(\delta\tau) \operatorname{sinc}(\delta f_d T) \cos(\delta\phi) + \eta_I, \\ Q &= \bar{A} D R(\delta\tau) \operatorname{sinc}(\delta f_d T) \sin(\delta\phi) + \eta_Q, \end{aligned} \quad (14.21)$$

where η_I and η_Q represent the noise in I and Q respectively, which are normalized by multiplication so that their noise power is given by

$$E(\eta_I^2) = E(\eta_Q^2) = 1, \quad (14.22)$$

where $E(x)$ denotes the expected value of x [14.2].

The early and late correlator concept is widely used in the BPSK code tracking due to the nonconvexity property of the triangular shape of code correlation function (Sect. 14.4.4 for more details including BOC signals). The correlator outputs, that is the baseband I/Q values at the early, prompt and late branches are six basic signal processing elements given as a function of signal parameter errors [14.1, 2]

$$\mathbf{B} = \mathbf{h}(D, d; \bar{A}, \delta\tau, \delta f_d, \delta\phi) + \boldsymbol{\eta} \quad (14.23)$$

$$\begin{aligned} \mathbf{B} &= [I_E, Q_E, I_P, Q_P, I_L, Q_L]^\top, \\ \mathbf{h} &= \begin{bmatrix} \bar{A} D R(\delta\tau - \frac{d}{2}) \operatorname{sinc}(\delta f_d T) \cos \delta\phi \\ \bar{A} D R(\delta\tau - \frac{d}{2}) \operatorname{sinc}(\delta f_d T) \sin \delta\phi \\ \bar{A} D R(\delta\tau) \operatorname{sinc}(\delta f_d T) \cos \delta\phi \\ \bar{A} D R(\delta\tau) \operatorname{sinc}(\delta f_d T) \sin \delta\phi \\ \bar{A} D R(\delta\tau + \frac{d}{2}) \operatorname{sinc}(\delta f_d T) \cos \delta\phi \\ \bar{A} D R(\delta\tau + \frac{d}{2}) \operatorname{sinc}(\delta f_d T) \sin \delta\phi \end{bmatrix}, \\ \boldsymbol{\eta} &= [\eta_{I_E}, \eta_{Q_E}, \eta_{I_P}, \eta_{Q_P}, \eta_{I_L}, \eta_{Q_L}]^\top, \end{aligned}$$

where d represents the early-to-late correlation spacing in chips (e.g., $d = 1$ for normal early-minus-late correlators).

Note that the above equation for six basic signal processing elements is for standard tracking of a BPSK signal. The number of early and late branches with a given correlator spacing is in general dependent on the modulation type of the signal (i.e., BOC), and the discriminator algorithms [14.3–6].

14.3 Signal Search and Acquisition

The signal search is the first phase of a GNSS receiver that determines if it receives signals from satellites at all. Due to the nature of the GNSS signals, the signal search includes a coarse estimation of the signal's Doppler and code phase. The signal search can intuitively be seen as a numerical evaluation of the signal's correlation function (14.18) in the two-dimensional Doppler and code phase space. If the peak magnitude of this function exceeds a certain threshold, then the signal is declared to be present and the position of the peak are the coarse estimates. This picture is not far from a fundamental theoretical concept, which is termed a generalized maximum likelihood ratio test [14.7].

14.3.1 Test Statistics

The basis for the acquisition logic are the correlator outputs (14.21). The correlator position in the code phase/Doppler domain with respect to the true values is denoted as $\delta\tau$ and δf_d . The carrier phase is not considered during signal acquisition and the total power S is given by

$$|S|^2 = I^2 + Q^2. \quad (14.24)$$

Under the assumption of a signal being present the total power S assumes the shape of a peak like the one shown

in Fig. 14.2 (or multiple peaks in the case of a BOC signal shown in Fig. 14.4), which has to be detected against a more or less uniform background of noise. The total power is written as

$$|S|^2 = A^2 R(\delta\tau)^2 \operatorname{sinc}(\delta f_d T)^2 + \text{noise}. \quad (14.25)$$

The expected value of the noise is easily evaluated, as terms linear in $\eta_{I,Q}$ vanish. Thus we have

$$E[\text{noise}] = E[\eta_I^2 + \eta_Q^2] = 2. \quad (14.26)$$

In the theory of signal detection, signal acquisition has to decide which of the following hypotheses is true:

- H_0 : Considered satellite signal is not present
- H_1 : Considered satellite signal is present.

Typically, signal acquisition is carried out for each satellite signal separately. The acquisition engine evaluates the total power S for a certain range of Doppler and code-phase values, searches the peak within this area and compares the value of the peak against a threshold γ . If the peak exceeds the threshold, the hypothesis H_1 is declared to be true, otherwise H_0 . To increase the sensitivity one can increase the *coherent* integration time T of (14.7), up to a limit determined by the computational resources (gate count in the case of a GNSS chip set or

central processing unit (CPU) load for software-based receivers), the oscillator stability, user dynamics or by the length of the broadcast navigation data bits/symbols or secondary codes. Usually it is in a range of 1–20 ms. A further sensitivity increase is achieved by a *noncoherent* integration; that is to compute the correlation function several ($= \nu$) times and to average them. Under some reasonable assumptions like constant Doppler and uniformly distributed carrier phase for each coherent integration interval, it can be shown mathematically that coherent plus noncoherent integration is in fact an optimal strategy [14.7]. In the following we assume that S_{nc} represents the coherent plus noncoherent integration result

$$S_{nc} = \sum_{n=1}^{\nu} |S_n|^2 \quad (14.27)$$

and that n denotes subsequent coherent integrations each over an interval of T . The total integration time is $T_{tot} = T\nu$.

Mathematically speaking, the function S_{nc} is the sum of 2ν squared zero-mean Gaussian random variables under the assumption H_0 , and S_{nc} follows a central chi-squared distribution $Q_{\chi^2, \alpha}$ with $\alpha = 2\nu$ degrees of freedom. The number of degrees of freedom 2ν is explained by recalling that each coherent integration has a contribution from the real I and the imaginary Q part. Using the definitions of [14.8] we have

$$P(S_{nc} > \gamma | H_0) = Q_{\chi^2, 2\nu}(\gamma). \quad (14.28)$$

The term $P(S_{nc} > \gamma | H_0)$ denotes the false alarm probability, that is the chance that the receiver incorrectly detects a signal even if none is present.

Under H_1 , the probability distribution of S_{nc} is a noncentral chi-squared distribution $Q_{\chi^2, \alpha; z}$ and the noncentrality parameter $z = 2\nu TC/N_0$ relates to the carrier-to-noise ratio C/N_0 and the number of noncoherent integrations ν

$$P(S_{nc} > \gamma | H_1) = Q_{\chi^2, 2\nu; 2\nu TC/N_0}(\gamma). \quad (14.29)$$

The term $P(S_{nc} > \gamma | H_1)$ is the detection probability, that is the ability of the receiver to detect signals that are actually present.

As an example the total power $\sqrt{S_{nc}}$ of a Galileo E1 composite binary offset carrier (CBOC) signal is shown in Fig. 14.4. Essentially the BOC(1,1) component of the CBOC signal determines the visual appearance, with the BOC(6,1) component being much smaller. The chosen acquisition settings $T = 8$ ms and $\nu = 5$ correspond to a medium to high sensitivity acquisition unit capable of detecting signals as low as approximately 30.5 dB-Hz as will be shown with equation (14.31). It can be seen that an unobstructed but low-elevation satellite signal has a clear peak, even showing the BOC(1,1) side lobes. Legacy GPS C/A receivers operating with $T = 1$ ms and $\nu = 1$ would not be able to detect this signal reliably and have difficulties acquiring signals near the horizon due to the diminished receiver antenna gain at low elevations.

Due to the squaring operation in the noncoherent integration, the noise floor gets positively biased. The ratio of the peak minus this bias divided by the standard

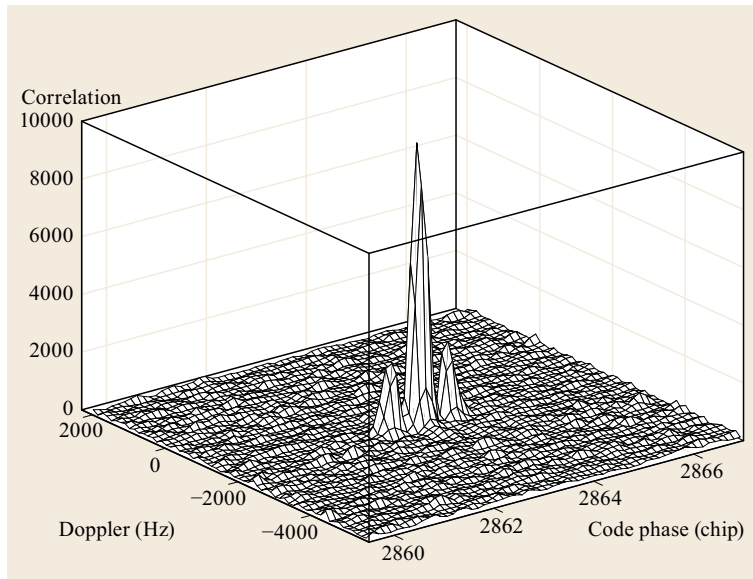


Fig. 14.4 Example of an E1C correlation function for a Galileo satellite at low elevation

deviation of the noise is called *signal-to-noise ratio* (SNR). For $\nu > 7-9$ the central limit theorem applies. Then the probability distributions can be approximated by Gaussian ones $\mathcal{Q}_{N;\mu;\sigma}$, with μ being the mean value and σ the standard deviation. Thus we can write

$$\begin{aligned} P(S_{nc} > \gamma | H_0) &\approx \mathcal{Q}_{N;2\nu;\sqrt{4\nu}}(\gamma), \\ P(S_{nc} > \gamma | H_1) &\approx \mathcal{Q}_{N;2\nu(1+TC/N_0);\sqrt{4\nu+8\nu TC/N_0}}(\gamma). \end{aligned} \quad (14.30)$$

With these approximations, we can establish a relationship between the carrier-to-noise ratio C/N_0 and the SNR value, which is useful to assess the sensitivity of an acquisition engine. As a first starting point, an SNR of 13 dB is required to detect a signal. A more precise description of the acquisition sensitivity will be given later. The SNR value derives from C/N_0 under these approximation as

$$\text{SNR} = \frac{2\nu TC/N_0}{\sqrt{4\nu}} = TC/N_0 \sqrt{\nu}. \quad (14.31)$$

One realizes that for large ν the SNR increase due to further noncoherent integration is twice as slow as for an increase in the coherent integration time T . This phenomenon is usually called *squaring loss*. For example, doubling the integration time increases the SNR value by 3 dB, and doubling the number of noncoherent integrations increases the SNR by 1.5 dB.

14.3.2 Acquisition Module Architecture

A block diagram of a modern GNSS acquisition engine is shown in Fig. 14.5. Acquisition is typically a process that takes place only when needed. In this case the IF sample selector cuts out a certain time span of GNSS signal samples from the received signal data stream. Two important preprocessing steps may take place, if the signal environment requires this.

For high-sensitivity acquisition, it may happen that strong and weak signals are received simultaneously.

For the example of an indoor receiver, one can think of a situation where a strong signal comes through the window and a weak signal has first to penetrate a wall before reception by the antenna. Unfortunately, the pseudo-random noise (PRN) codes from different satellites are not completely orthogonal; their cross correlation is slightly different from zero. Thus the correlation (total power) function S_{nc} not only contains the possibly present autocorrelation peak but also many cross-correlation peaks. The ratio between the autocorrelation peak and the largest cross-correlation peak under the assumption of equal power is called *cross-correlation protection*. It is a measure of how much signal power difference can be tolerated by the receiver still being able to differentiate the true autocorrelation peak from the unwanted cross-correlation peaks.

For the GPS C/A code it is around -22 dB. When the acquisition function S_{nc} of the weak signal is computed it might be contaminated by cross-correlation peaks from the strong signal (if the strong signal is e.g., 22 dB or more stronger than the weak signal). In that case it is beneficial to cancel (subtract) the strong signals from the incoming composite signal before computation of S_{nc} . This is possible with good precision, as a replica of a strong signal, whose parameters (code delay, Doppler, carrier phase and amplitude) can be accurately estimated, can be well reproduced inside the acquisition engine [14.9].

If the receiver is located inside a harsh RF environment (e.g., inside a mobile phone) a lot of interference, especially tone interference (isolated frequency peaks) is present. It might for example be caused by harmonics of the reference oscillator, local oscillators or any other clock signals inside the system. Those spikes in the spectrum have to be canceled by either fast Fourier transform (FFT) techniques or adaptive notch filters.

Finally, the computation of (14.27) takes place and a search and decision logic selects the correct peak. To gain some sensitivity and to increase the robustness,

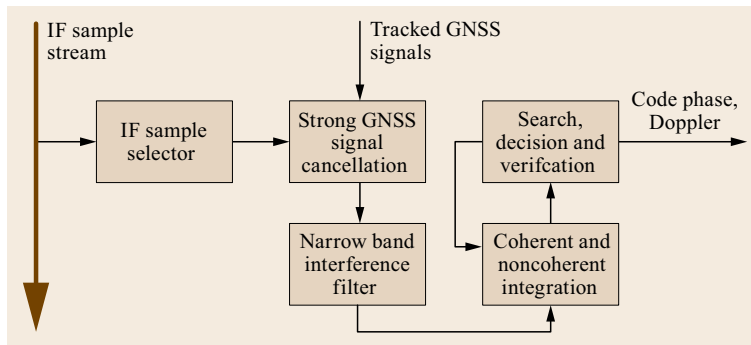


Fig. 14.5 Block diagram of a GNSS signal acquisition engine

a search and decision logic may select for example the ten highest peaks and then recompute (14.27) only for those code-phase and Doppler values, eventually using a larger number of noncoherent integrations. This is called *acquisition verification*. Verification is implicitly also done during the tracking process (a channel will not be able to track a false acquisition result for a longer time span), but verification in the tracking channels takes much more time and usually only a limited number of tracking channels are available.

14.3.3 Coherent Integration Methods

This section describes the computation of correlation values I and Q via numerical integration over the coherent integration time T , thus the evaluation of (14.18). This is the part of acquisition consuming most of the resources as it works with the native sample rate. In comparison, noncoherent integration or detection consume much fewer resources. The coherent integration is done by units called *correlators*.

The implementation of correlators for the acquisition unit differs from the tracking channel implementation as a wide Doppler range is considered, a wide code-phase range (typically the whole PRN code period) is considered and the accuracy requirements in terms of code-phase resolution or Doppler resolution are rather low. Thus different methods for code and Doppler correlation are considered.

Code Correlation

There are three main methods available to realize the coherent integration in code phase for acquisition purposes:

- Re-use of tracking channels for signal acquisition
- Parallel correlation with a matched filter structure
- Precorrelation FFT techniques.

The first method has historical importance or is used for special receivers focusing on a low-resource implementation. A tracking channel is programmed to continuously (but slowly) increase the code phase until the correlation peak is detected (obviously, the tracking channel has to be on the right Doppler frequency). This method is called *serial search* and may come along with Tong detectors [14.10] as a means to verify signal detection and to increase the sensitivity.

The second method is to set up a digital filter, whose filter coefficients are the PRN code values themselves (*matched filter*). The PRN code is resampled to the incoming sample rate. This filter is fed with the Doppler-compensated incoming signal. It can be

very efficiently realized in hardware using a pipelined structure [14.11]. The incoming signal is fed serially into the filter. Once initialized (i.e., after one PRN code period has been fed in), the structure outputs one value of the correlation function for each incoming sample. With each cycle, the output correlation function is shifted in code phase by one sample. The code-phase resolution in chips is given by the ratio of the PRN code rate divided by the sample rate.

The third method exploits the Fourier convolution theorem, which states that the computation of a correlation function in the time domain can be expressed as a multiplication in the frequency domain [14.7]. This drastically reduces the number of required operations and is the method of choice if a software-based solution is targeted. Various methods like zero padding or resampling exist to cope with the fact that an FFT approach intrinsically works with periodic signals [14.7]. The code-phase resolution in chips is again given by the ratio of the PRN code rate divided by the sample rate.

Doppler Correlation

The code correlation methods from before have to be carried out after Doppler removal. In the most simple way, a carrier NCO is used for this purpose. The term *carrier NCO* is used to denote the responsible unit to generate a sine/cosine wave with a certain phase and frequency. The received signal samples are multiplied with carrier NCO samples (sine and cosine part) and the code replica samples before carrying out the coherent integration. If the selected carrier NCO frequency matches the true Doppler frequency, the correlation value becomes a maximum; the Doppler has been removed.

A certain Doppler search strategy has to be employed to change the carrier NCO frequency in a proper way, each time one PRN code sequence has been searched. This is done until the signal is detected. To give an example, the receiver may start with the lowest possible frequency of for example -6 kHz and gradually increases the frequency in steps of 500 Hz until the highest possible frequency of for example 6 kHz is reached. This approach can be slow if the Doppler search range is wide, but it is suitable if the Doppler is approximately known (e.g., in case of reacquiring a signal).

A more efficient way to search a wide Doppler range is to use *postcorrelation FFT* techniques [14.12]. They exploit the fact that the multiplication with the carrier NCO together with the integration resemble the definition of the discrete Fourier transform. If the integration over M samples is subdivided into N chunks,

then one may write

$$\begin{aligned}
 & \sum_{k=1}^M s_k \exp\left(2\pi j \hat{f}_d T_s k\right) \\
 &= \sum_{n=1}^N \sum_{k=(n-1)\frac{M}{N}+1}^{\frac{nM}{N}} s_k \exp\left(2\pi j \hat{f}_d T_s k\right) \\
 &= \sum_{n=1}^N \left[\exp\left(2\pi j \hat{f}_d T_s (n-1)\frac{M}{N}\right) \right. \\
 &\quad \times \left. \sum_{k=1}^{\frac{M}{N}} s_{k+(n-1)\frac{M}{N}} \exp\left(2\pi j \hat{f}_d T_s k\right) \right] \\
 &\approx \sum_{n=1}^N \left[\exp\left(2\pi j \hat{f}_d T_s (n-1)\frac{M}{N}\right) S_n \right]. \quad (14.32)
 \end{aligned}$$

Here $s_k = r(k)C(T_s k - \hat{\tau})$ is a placeholder for the received signal multiplied with the code replica and

$$S_n = \sum_{k=1}^{\frac{M}{N}} s_{k+(n-1)\frac{M}{N}} \exp\left(2\pi j \tilde{f}_d T_s k\right). \quad (14.33)$$

The short coherent integrations S_n are performed all with the same carrier NCO frequency \hat{f}_d being an integer multiple of $N/(MT_s)$. For example, if a coherent integration time of $MT_s = 4$ ms is targeted, one can compute eight correlation functions S_n , each based on $NT_s = 0.5$ ms of a GNSS signal assuming $\tilde{f}_d = 0$ Hz. Applying for each code phase bin separately an eight-point FFT gives the 4 ms correlation function for the Doppler values of $\hat{f}_d = -250$ Hz, -187.5 Hz, \dots , 187.5 Hz. This is computationally much more efficient than doing a full 4 ms correlation for all eight Doppler bins. There is a certain correlation loss L_{PC} involved depending on the difference of \hat{f}_d and \tilde{f}_d , evaluated as

$$L_{PC} = 20 \log_{10} \left| \text{sinc} \left[NT_s (\hat{f}_d - \tilde{f}_d) \right] \right|, \quad (14.34)$$

which assumes in this example for a maximum difference of 250 Hz a value of -3.96 dB. To which extend this loss can be tolerated is application dependent.

In the case where the code correlation is performed with FFT methods, a very efficient way exists to search different Doppler bins. Before multiplying the Fourier transform of the received signal with the transformed replica signal, the transformed replica signal is circularly shifted by a certain number of samples. A shift of one sample corresponds to a Doppler step of $(LT_s)^{-1}$ Hz with L being the length of the FFT. Thus all Doppler bins can be searched and only the multiplication and the

inverse FFT (but not the forward FFTs) have to be carried out for each Doppler bin.

It should be mentioned that a number of further advanced FFT algorithms have been developed mostly by David Akopian, which are summarized in [14.7] for example. They include a lossless postcorrelation FFT method as well as an highly efficient algorithm applicable if the code search range can be limited. The use of Fourier methods in the one or other form within an acquisition unit is inevitable if numerically optimal performance is sought.

14.3.4 Search Space

Acquisition is a two-dimensional search in the code phase/Doppler dimension. A reduction of the search space not only saves computational resources but also increases the sensitivity. A suitable resolution for the code phase and Doppler grid has to be chosen.

Code Direction

The search in code direction typically covers the full PRN code period because a precise time with an accuracy of better than the PRN code period (e.g., 1 ms) is usually not available at startup. Furthermore, matched filter techniques or FFT techniques prefer to work with the whole PRN code sequence. Two notable exceptions from this rule should be mentioned.

The first case considers encrypted GNSS signals. Typically they are based on very long PRN codes, which provide some protection against PRN code decoding and dissemination. For example, the GPS P(Y)-code is one week long and cannot be acquired in this way. If direct P(Y)-code acquisition is targeted (and not a handover from the GPS C/A code), then a smaller fraction of the P(Y)-code is search in larger piece of incoming signal (or vice versa). The length of the larger portion has to reflect the receiver uncertainty of its prediction of the code phase.

Second, modern GNSS signals typically consist of a primary and a secondary code. The direct acquisition of the combined code (which has a typical length of 10–100 ms) is usually difficult due to computational limitations but mostly due to user dynamics and receiver clock instability. Instead one considers only the primary code. The secondary code can be regarded as a navigation data bit/symbol.

The code phase resolution has to be high enough in order not to miss the correlation peak. In the worst case, the peak is right in the middle of two correlation values. For a BPSK signal, a minimum of $r = 2$ samples per chip shall be used, while for a BOC(n, m) signal a number of $r = 4n/m$ samples per chip is required. This is illustrated with Fig. 14.6 where the worst case of maxi-

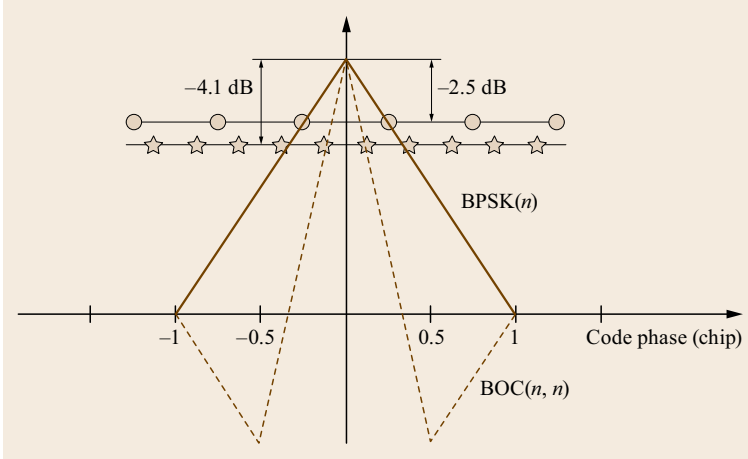


Fig. 14.6 Impact of finite code resolution for a BPSK(n) and a BOC(n, n) signal using 2 and 4 samples per chip, respectively

imum code phase mismatch for the suggested resolution settings is shown. The figure shows one BPSK and one BOC correlation function with the main peak just between two code phase bins denoted as circles for the BPSK signal and stars for the BOC signal.

For a BPSK signal, the maximum loss is then

$$20 \log_{10} \left(1 - \frac{1}{2r} \right) = -2.5 \text{ dB} . \quad (14.35)$$

For BOC(n, m) signals, it is

$$20 \log_{10} \left(1 - \frac{1}{2r} \frac{2n+m}{m} \right), \quad (14.36)$$

which evaluates to -4.1 dB for a BOC(n, n) signal. These are worst-case losses and average mismatch losses are smaller. If they are too high for a certain application, then the resolution r needs to be increased at the cost of higher computational demands.

Doppler Direction

The Doppler search range has to be wide enough to cover the user velocity, satellite velocity and the receiver clock drift. On the L1 frequency it is on the order of ± 6 kHz for a static receiver with an oscillator having at least temperature-compensated crystal oscillator (TCXO) quality (expressed as an absolute frequency stability over the operating temperature range of 1 ppm).

Once the receiver clock drift has been estimated, it usually remains stable for hours (if the operating temperature is sufficiently stable). If the user position and velocity is approximately known and an almanac is available, the Doppler search range can be narrowed down to the order of ± 100 Hz.

If the clock drift is unknown, it can be coarsely estimated from a single first satellite signal, provided that

an approximate position plus velocity and an almanac are available. Only for this first signal, the full Doppler search range has to be employed. The following signals can then be found in a narrow Doppler range.

The correlation value of (14.25) depends on a sinc function of the Doppler frequency error as drawn in Fig. 14.7. Thus the Doppler resolution should be at least $\Delta f_d = 1/T$ (shown as circles in Fig. 14.7), which gives a maximum loss of

$$20 \log_{10} \text{sinc} \frac{\Delta f_d T}{2} = -3.9 \text{ dB} . \quad (14.37)$$

A finer resolution is preferable if the required computational resources are available.

14.3.5 Acquisition Performance

The acquisition performance is expressed in signal processing terms as the probability of detection (for a given signal power and false alarm rate) and the time needed

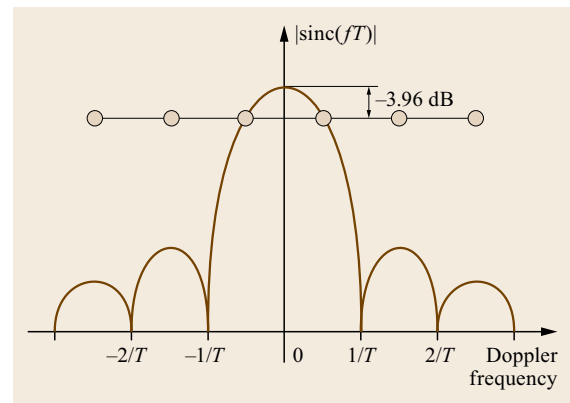


Fig. 14.7 Impact of a finite Doppler resolution during acquisition. Circles are spaced at $\Delta f_d = 1/T$

to perform the operation. The time needed contributes to the time-to-first-fix (TTFF) and depends on the number of correlation values to be computed.

As the search space of Sect. 14.3.4 usually includes many, let's say N_G , bins, it is useful to distinguish single-bin and system probabilities. Single-bin probabilities are used in (14.28)–(14.30) and refer to the case that only a single correlation value is considered (either the signal is present in this test bin or not). System probabilities P' consider the whole search space instead. According to [14.13] they are related to the single-bin probabilities by the following equations

$$\begin{aligned} P'(S_{nc} > \gamma | H_0) &\approx 1 - [1 - P(S_{nc} > \gamma | H_0)]^{N_{uc}}, \\ P'(S_{nc} > \gamma | H_1) &\approx P(S_{nc} > \gamma | H_1). \end{aligned} \quad (14.38)$$

To discuss system probabilities, the concept of the number of uncorrelated search bins N_{uc} is important. Usually the correlation values are statistically correlated, especially if a fine code phase of Doppler resolution is used. Only if two test bins are separated far enough they can be considered as statistically uncorrelated. N_{uc} is approximately given as

$$N_{uc} = \Delta T f_c T \Delta f_d, \quad (14.39)$$

where ΔT is the search range in time (code phase direction) in seconds, f_c is the chipping rate of the PRN code in chips per second, T is the coherent integration time in seconds and Δf_d is the Doppler range in Hz.

Choosing a proper system false alarm rate $P'(S_{nc} > \gamma | H_0)$ to achieve a minimum detection probability $P'(S_{nc} > \gamma | H_1)$ for a given signal power is highly application dependent. Both are linked to each other via the threshold γ and cannot be optimized simultaneously.

If we assume that the search space contains N_G grid points (usually $N_G > N_{uc}$) and the system needs D_W seconds to compute M_G values of S_{nc} , then the whole grid is searched in $D_W N_G / M_G$ seconds.

One may, however, decide to perform an acquisition test $S_{nc} > \gamma$ every time the M_G values are available. Then the mean time T_M needed to acquire a single GNSS signal can be computed under several assumptions as [14.13]

$$T_M = \left(\frac{2 - p_d}{2p_d} \right) \left(\frac{k [1 - (1 - p_{fa})^{M_G}] + 1}{M_G} \right) N_G D_W. \quad (14.40)$$

The definitions $p_d = P(S_{nc} > \gamma | H_1)$ and $p_{fa} = P(S > \gamma | H_0)$ are used. The symbol k is a penalty factor used to characterize the time needed to detect a false acquisition.

If a false signal is incorrectly detected, the system is assumed to spend kD_W seconds to detect its error and then to continue with the normal acquisition procedure. We assume that the true signal parameters lie exactly on one of the grid points and that the correlation values for all other grid points follow a H_0 distribution. In other words, if the signal is present on one grid point, it is not present on all the others. We also assume that the position of the true grid point is uniformly distributed over all grid points. Finally, for (14.40) it is assumed that the correlation values on different grid points are uncorrelated, that is $N_G = N_{uc}$.

14.3.6 Handling Data Bits and Secondary Codes

Possibly present data bits or data symbols and secondary codes represent a certain challenge during signal acquisition. Those data bits and their exact transition times are generally unknown to the acquisition engine. Whereas the secondary code is known, the code phase within the secondary code is generally unknown during acquisition. Thus the secondary code has a similar effect on acquisition as an unknown data bit. In the following we use the term data bit but do mean data bits (or data symbols in the case where a forward error correction scheme is used) or secondary code chips.

One may easily envisage a situation where a data bit transition occurs in the middle of a coherent integration interval and the first and second half of the correlation cancel each other.

One method, which is called the half-bit method, to cope with this unknown transitions is to subdivide the signal into two interleaving sequences of segments with each of the duration of half the data bit. One of the sequences is then free of transitions. The acquisition engine works on both sequences thereby doubling the number of search bins. It will detect the signal only in one sequence.

Data bit transitions correlate with a residual Doppler frequency within the received navigation signal. For example, a data bit transition in the middle of the integration interval and a residual Doppler frequency of $1/(2T)$ nearly cancel each other. This effect can be exploited especially if a large number of non-coherent integrations is used. Then one can show that the average loss induced by randomly occurring data bit transitions is -2.14 dB for a data bit length of 20 ms and a coherent integration time of 16 ms [14.9]. This loss is an average value over all Doppler frequencies and over the whole primary code-phase range. Although the loss for individual bins might be higher, simply ignoring those data bit transitions is a viable approach during acquisition.

14.4 Signal Tracking

After the coarse estimate of initial code delay and carrier Doppler by the acquisition block, the signal tracking is performed to obtain fine estimates of signal parameters of interest. A number of traditional signal tracking loop architectures such as phase-locked-loop (PLL) for carrier-phase tracking, frequency-locked-loop (FLL) for carrier Doppler frequency shift tracking, and delay-locked-loop (DLL) for code delay tracking are widely used as engineering standards in modern digital GNSS receivers.

14.4.1 Architecture

In fact, the main purpose of a signal tracking loop is to adjust the input of the local replica signal generators to match the received signals. This is done by adjusting the input of the NCO, which is the rate of change of the signal parameter of interest. This means that the tracking loop is implemented in a form of feedback systems that are closed by the NCO. The signal tracking loop can be considered as the special case of an output feedback control system that has an integrator in the feedback loop as an actuator to track the input signal. Thus, the resulting tracking loop filter is a type of proportional and integrate (PI) controller in control theory. In a PLL (or FLL) the output of the discriminator is the phase (or frequency) error estimate of the input

carrier signal, while the code delay error estimate is for a DLL.

Figure 14.8 shows a high-level block diagram of a single-channel signal tracking engine in typical digital GNSS receivers.

The single tracking loop in a single channel of a GNSS receiver is composed of correlators, discriminators, loop filters and NCOs for the code and carrier tracking loops, where the digitized IF signal sequences are applied to the input. The programmable designs of the integration time interval in correlators, the type of discriminators as well as the order and bandwidth of loop filters determine characteristics of the signal tracking loop in terms of the steady-state error and the robustness to dynamic stress [14.10]. The carrier Doppler aiding to code tracking and/or external velocity aiding to the carrier tracking loop can be employed to enhance the signal tracking performance. The lower feedback in Fig. 14.8 illustrates the carrier tracking loop in a traditional digital GNSS receiver, while the upper part is the code tracking loop.

The operation of the signal tracking engine is as follows. The carriers in the digital IF signal sequences are wiped off by the replica carrier signals to produce the I and Q signal components. The replica carrier signals are synthesized by the carrier generator for which the input is the carrier NCO output. The I and Q signal compo-

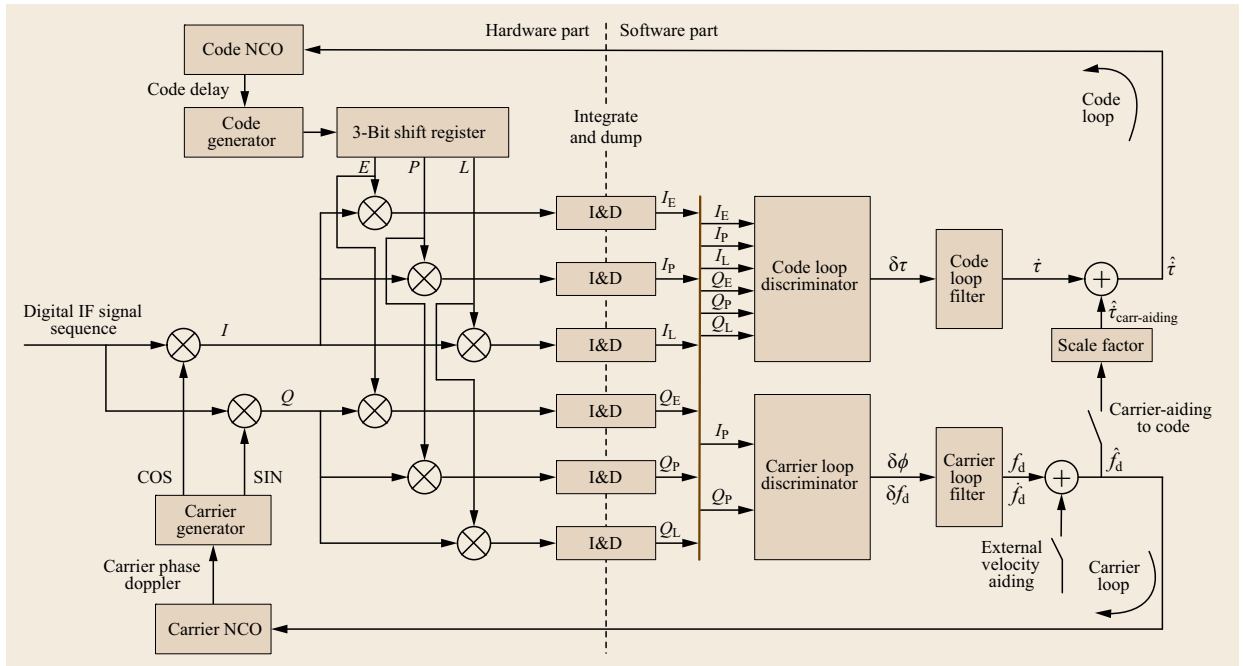


Fig. 14.8 Block diagram of a GNSS signal tracking engine

nents are then correlated with the replica codes at early, prompt, and late branches (for the most simple case of standard tracking of a BPSK signal). They are, similar to the previous case, synthesized by the code generator with a 3 bit shift register for which the input is the code NCO output. In the closed loop operation, the carrier and code NCOs are controlled by the carrier and code tracking loop, respectively.

Normally, the correlator output at prompt branches is used in the carrier tracking whereas the correlator output at early and late branches are used in the code tracking. In particular, the arctangent of Q/I is used in the carrier discriminator to determine the carrier angle (i.e., phase $\delta\phi$) of the incoming signal with respect to the in-phase (I) replica signal of the prompt. For code tracking, the combination of outputs at early and late branches (e.g., early-minus-late) is used in the code discriminator to estimate the delay $\delta\tau$ of the incoming codes with respect to the replica codes. Therefore, the absolute value of I at the prompt branch should be maximum (for PLL), and the vector sum of I and Q at early and late branches should be balanced (for DLL) when the replica signal is successfully aligned to the incoming signal. Finally, the sign of I represents the navigation data bit/symbol.

14.4.2 Tracking Loop Model

A basic model for a single-channel signal tracking loop in a GNSS receiver is shown in Fig. 14.9. Note that the input of the signal tracking loop is the signal (i.e., wave) in the continuous time domain, $r(t; \theta(t))$, which is nonlinear in the signal parameter of interest, θ , and its output is the corresponding estimated signal, $\hat{r}(t; \hat{\theta}(t))$. Depending on the type of loop, the parameter θ may either be the code phase τ (DLL), the Doppler shift f_d (FLL), or the carrier phase ϕ (PLL).

The discriminator output is the signal parameter error, $\delta\theta$, obtained by comparing the incoming signal and the local replica signal. In fact, this process is accomplished by the combination of the correlator outputs, that is the baseband I and Q signal components computed by the nonlinear operation with the received signals and the locally generated replica sig-

nals as in (14.23). This discriminator output contains noise that should be efficiently filtered out by the loop filter. Therefore, the tracking process is sensitive in terms of noise effects and dynamic responses rather than the coarse acquisition process. Note that the output of the loop filter is the rate of change information of the signal parameter of interest, $\dot{\theta}$, which is then integrated in the NCO to predict the signal parameter estimate, $\hat{\theta}$, for the next step. This signal parameter estimate is used in the local signal generator to produce the estimated local replica signals for the correlation.

In order to obtain a mathematical description of the tracking loop suitable for subsequent design and analysis, it is desirable to approximate linearly the nonlinear operation of the actual tracking loop. Under the assumption that the tracking error is small enough and that the input noise is uncorrelated with the locally generated replica signal, the linearized tracking loop in the Laplace domain is obtained for actual input signal parameters as shown in Fig. 14.10, where $s = \sigma + j\omega$ denotes the Laplace operator. The difference between Fig. 14.10 and Fig. 14.9 is that the nonlinear operation of the correlator and discriminator for the incoming and local replica signals is substituted by a simple comparator with the linear dependence on signal parameters, where $\theta(s)$ and $\hat{\theta}(s)$ represent the input of the tracking loop and the output from the NCO.

In this linear model we assume that the comparator measures the difference of these two signal parameters to provide the input signal $\delta\theta$ to the loop filter. Assum-

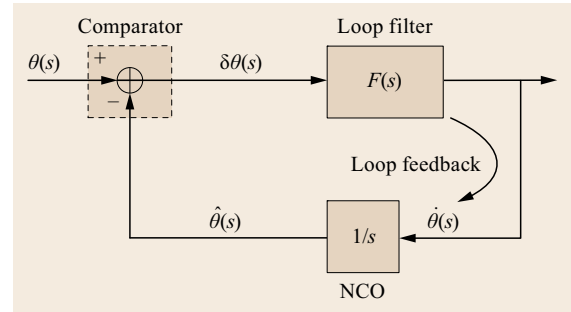


Fig. 14.10 Linear model of a signal tracking loop in the Laplace domain

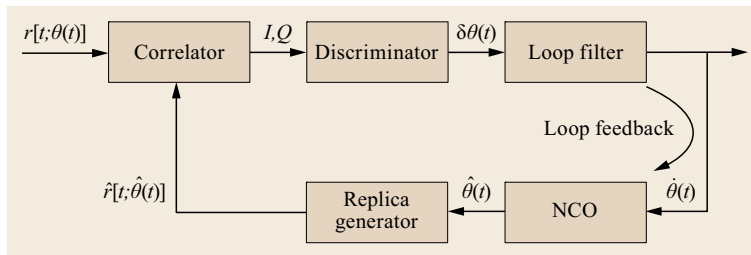


Fig. 14.9 Nonlinear model of a signal tracking loop in the time domain

ing noise contaminated signal parameters, the output of the comparator is modeled by

$$\delta\theta(s) = \theta(s) - \hat{\theta}(s), \quad (14.41)$$

where $\theta(s) = \theta_0(s) + n_\theta(s)$ is the sum of the nominal value θ_0 of θ and the corresponding noise n_θ .

The loop filter transfer function $F(s)$ is defined as

$$F(s) = \frac{\dot{\theta}(s)}{\delta\theta(s)}, \quad (14.42)$$

where $\dot{\theta}(s)$ and $\delta\theta(s)$ represent the Laplace transforms of the input $\dot{\theta}(t)$ and the output $\delta\theta(t)$ respectively.

The NCO is a simple integrator, therefore its transfer function $N(s) = 1/s$, and its output is modeled by

$$\hat{\theta}(s) = \frac{1}{s} \dot{\theta}(s). \quad (14.43)$$

Assuming a unity gain for the discriminator and NCO, the overall loop transfer function $H(s)$ is given by

$$H(s) = \frac{\hat{\theta}(s)}{\theta(s)} = \frac{N(s)F(s)}{1 + N(s)F(s)} = \frac{F(s)}{s + F(s)} \quad (14.44)$$

and its error transfer function $H_e(s)$ is also given by

$$H_e(s) = 1 - H(s) = \frac{s}{s + F(s)}. \quad (14.45)$$

Then, the input signal $U(s)$ (e.g., a unit step, ramp, acceleration, etc.) generates the error function given by

$$\varepsilon(s) = H_e(s)U(s). \quad (14.46)$$

The steady-state error can be found by applying the final value theorem of the Laplace transform to (14.46) given by

$$\varepsilon(s=0) = \varepsilon(t=\infty) = \frac{d^n G}{d^n} \omega_0^n, \quad (14.47)$$

where n is the loop filter order and $d^n G/dt^n$ is the maximum line-of-sight dynamics. If, for example, $n = 2$, then $d^2 G/dt^2$ represent the line-of-sight acceleration. For a n th order tracking loop, the natural frequency is denoted by ω_0 and can be computed by using (14.48).

Note that the above steady-state error is a function of the loop filter order n ; therefore, the type of steady-state error for the signal tracking loop is dependent on the loop filter order.

The equivalent noise bandwidth of $H(s)$ is defined as

$$B_n = \int_0^\infty |H(j\omega)|^2 d\omega \approx \begin{cases} \frac{\omega_0}{4} & n = 1, \\ \frac{\omega_0}{1.89} & n = 2, \\ \frac{\omega_0}{1.2} & n = 3, \end{cases} \quad (14.48)$$

where ω is the angular frequency (rad).

14.4.3 Correlators

The correlation process (also known as predetection integration process) in (14.18) is performed by correlators. This is done by two steps:

- The integration and dump for a given interval at hardware correlators after mixing (i. e., multiplying) incoming and local replica signals
- The accumulation of the integration and dumped values for the next N samples in the software.

In this way the receiver's signal processor can extend the integration time with a minimum on hardware resources; for example a 1 ms integration and dump is done with hardware correlators and the accumulation of the next 20 samples for 20 ms of integration time (GPS C/A code) is done in the software. The integration and dump interval, which is normally set to be at least equal to or larger than a single code epoch interval in the tracking loops, multiplied by the N accumulation time in software, determines the bandwidth of the predetection process. The resultant integration time should be as long as possible for the high sensitivity in weak or interfered signals and as short as possible for the robustness to high signal dynamics.

The resulting outputs of correlators are six basic signal processing elements as in (14.23), in particular for BPSK-like signals, which are inputted into the discriminator to produce the signal parameter errors. Depending on the code type and the code tracking algorithm, a larger number of correlator channels can be used in addition to the six basic signal processing elements.

14.4.4 Discriminators

The purpose of a discriminator is to extract the signal parameter error information from the correlator outputs I and Q at the early, prompt, and late branches, which are nonlinear functions of the signal parameters of interest to be estimated as in (14.23). The output of the nonlinear discriminator function is the combination of the correlator outputs that is linearly increasing with the input signal parameter error within the linear region, such that the discriminator senses the amount of error in the replica signal with respect to the incoming signal (so-called S -curve). The output of the discriminator is then provided to the NCO for the next step in a negative feedback to control the code/carrier generators. This discriminator output may contain a greater noise effect, therefore the loop filter are employed before the NCO in the feedback in order to efficiently suppress the noise effect.

The type of discriminator algorithm determines the type of tracking loop (i. e., PLL, FLL or DLL) as well as relevant characteristics (i. e., phase reversal sensitivity, signal amplitude sensitivity, computational load, pull-in-range, etc.).

PLL and FLL Discriminators

The carrier loop discriminator determines characteristics of the carrier tracking loop as a carrier-phase tracking loop or a carrier Doppler tracking loop.

For the carrier-phase tracking loop, there are mainly two types: a pure PLL and a Costas PLL. The pure PLL is sensitive to the presence of bit/symbol modulation on the signal. The Costas PLL is insensitive to that if the integration time of the correlator to produce the baseband I and Q does not straddle the data bit/symbol transitions. Both discriminators produce the carrier-phase error (but with different pull-in-ranges) that are used as input to the PLL filter to reduce the noise.

For the carrier Doppler tracking loop, the FLL discriminator produces the carrier Doppler frequency errors that are also used as input into the FLL filter for the same reason.

The carrier-phase tracking loops are more accurate but more sensitive to dynamic stress than the carrier Doppler tracking loop. There is a trade-off between these three methods as shown in Table 14.1. For better accuracy, a PLL for pilot channel or a Costas loop for data channel with lower order and narrow bandwidths should be employed, and vice versa for better robustness to dynamics.

The basic idea of the carrier loop discriminator starts from Q/I at the prompt branch in (14.23) to extract the signal parameter error information (i. e., the carrier-phase error). Under the assumption of no noise in (14.23), the carrier-phase error can be obtained by

$$\frac{Q}{I} = \frac{\sin(\delta\phi)}{\cos(\delta\phi)} = \tan(\delta\phi) \approx \delta\phi \quad \text{for } \delta\phi \approx 0. \quad (14.49)$$

Taking the arctangent in both sides, we obtain

$$\tan^{-1}\left(\frac{Q}{I}\right) = \delta\phi. \quad (14.50)$$

For pure PLLs, the arctangent in (14.50) is substituted by a four-quadrant arctangent in order to remain linear over the full input error range of $\pm 180^\circ$, whereas arctangent for Costas PLLs remains linear over half of the input error range ($\pm 90^\circ$) [14.10]. There are other similar PLL discriminator functions with less computational burden, such as $\text{sign}(I)Q \approx \delta\phi$ or $IQ \approx 2\delta\phi$,

Table 14.1 Trade-off in GNSS receiver carrier tracking loop design

Loop property	For robustness to dynamics	For accuracy
Carrier tracking loop type	FLL	PLL (for pilot channel) or Costas (for data channel)
Loop filter order	High	Low
Loop noise bandwidth	Wide	Narrow

but the slopes of the discriminator outputs are proportional to signal amplitude, which causes nonoptimal properties depending on the signal-to-noise ratio. Beside the algorithm described before, there are several variants of PLL and FLL discriminator algorithms that produce slightly different characteristics in terms of coherent/noncoherent, normalization, and computational complexity [14.10].

By applying I to a deadbeat detector (like a signum function), we can obtain the navigation message bits/symbols. However, due to the $\pm 180^\circ$ of phase reversal in Costas PLLs, the ambiguity in the detected navigation symbols exists. This is solved later in the frame synchronization process by using the known preamble bits at the beginning of each subframe (Sect. 14.5.3).

The pull-in range of a PLL is the set of all the initial conditions that lead to phase lock, provided that a stable singular point exists [14.14]. Within this pull-in range the discriminator output is approximately linearly proportional to the input error, and the loop filter can operate correctly to reduce the tracking error finally reaching the zero-crossing point. Therefore, the input value to the loop filter at the initial stage should be sufficiently converged within this pull-in range for the correct loop filter operation; that is the acquisition output (e.g., the coarse code delay and Doppler) should be sufficiently accurate. If this cannot be achieved, the channel has to be operated in FLL mode until the Doppler error gets small enough. Additionally, this pull-in range is significantly related to the dynamic range of the loop filter.

In order to derive the FLL discriminators, we start from the definition of the Doppler frequency, which is the difference between two adjacent phases of baseband samples within the same data symbol [14.15]

$$f_d \equiv \frac{(\phi_{k+1} - \phi_k)}{T}, \quad (14.51)$$

so that the Doppler error, the output of the FLL discriminator, is obtained by applying δ to both sides

$$\delta f_d \equiv \frac{(\delta\phi_{k+1} - \delta\phi_k)}{T}. \quad (14.52)$$

Here k is used to index baseband samples and T is the coherent integration time.

Applying the tangent function to the right-hand side of (14.52) and using the relationship between $\delta\phi$ and the baseband I and Q signal components in (14.50), we obtain

$$\begin{aligned}\tan(\delta\phi_{k+1} - \delta\phi_k) &= \frac{\tan \delta\phi_{k+1} - \tan \delta\phi_k}{1 + \tan \delta\phi_{k+1} \tan \delta\phi_k} \\ &= \frac{\frac{Q_{k+1}}{I_{k+1}} - \frac{Q_k}{I_k}}{1 + \left(\frac{Q_{k+1}}{I_{k+1}}\right) \left(\frac{Q_k}{I_k}\right)} \\ &= \frac{I_k Q_{k+1} - I_{k+1} Q_k}{I_k I_{k+1} + Q_k Q_{k+1}} \\ &= \frac{\text{cross}}{\text{dot}}\end{aligned}\quad (14.53)$$

with

$$\begin{aligned}\text{cross} &= I_k Q_{k+1} - I_{k+1} Q_k, \\ \text{dot} &= I_k I_{k+1} + Q_k Q_{k+1}.\end{aligned}$$

Finally, the combination of dot and cross divided by T represents the Doppler frequency error

$$\frac{\tan^{-1} \left(\frac{\text{cross}}{\text{dot}} \right)}{T} = \delta f_d. \quad (14.54)$$

Intrinsically, the FLL tracks the carrier frequency and not the carrier phase so that the FLL discriminator is insensitive to $\pm 180^\circ$ phase reversals but its integration time must not straddle the navigation data bit/symbol transitions. If not, the phase jump caused by 180° phase reversals acts as a high-frequency component that affects the rest of the processing. Beside the algorithm described before, there are several variants of FLL discriminator algorithms that produce slightly different characteristics in terms of optimality depending on the signal-to-noise ratio, the pull-in range corresponding to the integration time and computational complexity [14.10].

Note that the pull-in range of the FLL discriminator is inversely proportional to the integration time because of the fact that the FLL discriminator output is obtained by using two consecutive pairs of carrier-phase discriminator outputs divided by the integration time.

DLL Discriminator

The basic idea of the code loop discriminator is almost the same as the PLL discriminator, but it uses the early and late branches rather than the prompt branch of the carrier loop. In fact, the PLL discriminator is obtained from the first- and/or second-order derivative of the maximum likelihood estimation (MLE) cost

function, while the cost function for the code delay, which is a triangular shape of correlation function, is not differentiable at the peak when no error exists. Therefore, half-chip early and half-chip late sampling of the triangular correlation function divided by the given sampling interval, which is an approximate first-order derivative of the cost function, is used.

Figure 14.11 shows how the early, prompt, and late correlators change as the offset of the locally generated code replicas are advanced with respect to the incoming satellite's code signal and the corresponding normalized early-minus-late discriminator output for the relevant five cases of replica code offsets.

If the replica code is aligned, then the early and late branches are equal in amplitude and no error is generated by the discriminator. If not, the early and late samples are not equal by an amount proportional to the code offset, so that the DLL discriminator outputs the relevant error, which is then fed into the closed loop operation through the loop filter to filter out the noise effect.

There are several variants of DLL discriminator algorithms that produce slightly different characteristics in terms of required coherent-noncoherent integrations, accuracy, availability depending on the carrier-lock condition, correlator spacing and computational complexity, described in [14.10]. The most common one is a normalized early-power minus late-power code discriminator

$$\delta\tau = \frac{\alpha_d}{2(2-d)} \frac{I_E^2 + Q_E^2 - I_L^2 - Q_L^2}{I_P^2 + Q_P^2}, \quad (14.55)$$

which is in first-order linearly proportional to the code tracking error. The constant α_d has to be chosen depending on the modulation scheme to achieve unity slope.

In addition, due to the approximation of the non-convexity cost function by the early and late correlator concept, the correlation spacing d plays an important role in the DLL performance. For example, narrowing the correlator spacing d is beneficial in the reduction of tracking errors in the presence of both noise and multipath whereas a wider precorrelation bandwidth is required, coupled with higher sampling rates and higher processing speed as well as losing the robustness to dynamics [14.16].

14.4.5 Loop Filters

One of the most important factors in the signal tracking software is to obtain a loop filter, $F(s)$ in (14.42). Since the discriminator output contains much noise, the purpose of the loop filter is to reduce noise in order to produce an accurate estimate of the original signal.

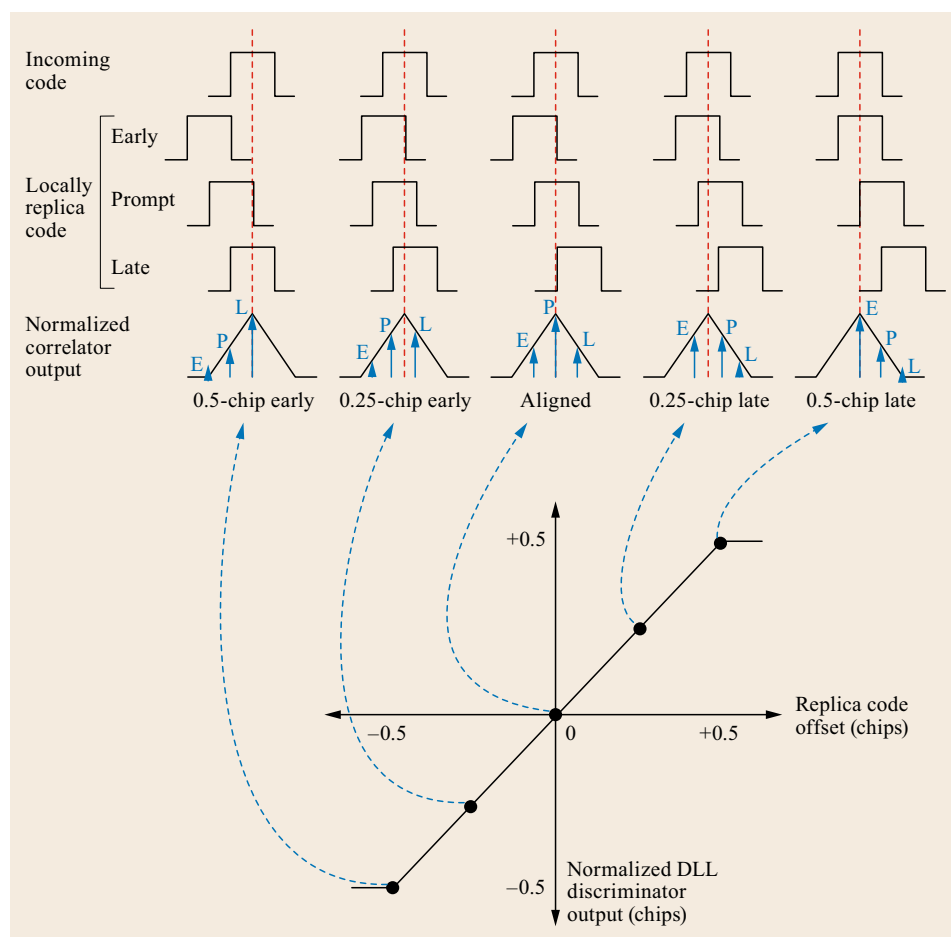


Fig. 14.11 Early-minus-late DLL discriminator (after [14.10])

The design criteria for the loop filter is to design an optimal filter to minimize the root mean square (RMS) noise error based on the constraint of a transient error specification. This means that the RMS error due to the noise interference in steady state should be minimized in terms of the NCO (code or carrier) phase jitter, while the transient error in the signal (code or carrier) phase due to dynamics should also be maintained at a specific amount at the same time.

There are many different approaches to designing a digital loop filter. One of the design approaches, which is widely used, is to design the overall signal tracking loop in the continuous time domain with existing knowledge of analog loop filters, and then to implement this in the discrete time domain.

One solution for this is offered by the variational method for optimization via a Lagrangian multiplier. An optimal solution for the loop filter in the continuous time domain is given by a function of the loop noise bandwidth as in (14.48), which limits the noise in

the loop and can be interpreted as a normal filter coefficient such as a decay ratio, damping ratio, or natural frequency [14.17].

Figure 14.12 shows block diagrams of first-, second- and third-order loop filters. Note that an integrator ($= 1/s$) for the NCO is located additionally in the open loop; the order of a loop filter determines the order of $H(s)$, not $F(s)$, that is for an n th order tracking loop the denominator of the transfer function $H(s)$ is an n th order function of s .

The signal tracking loop obtained in the continuous time domain as a differential equation is then transformed into the corresponding discrete equivalent in a difference equation by applying the digital integration method (also known as the z -transform method). This is done by replacing s in the continuous transfer function with a function of z .

Under the assumption of a simple integration method, the closed-loop transfer function, $H(s)$, can be represented in state-space form for any order of loop

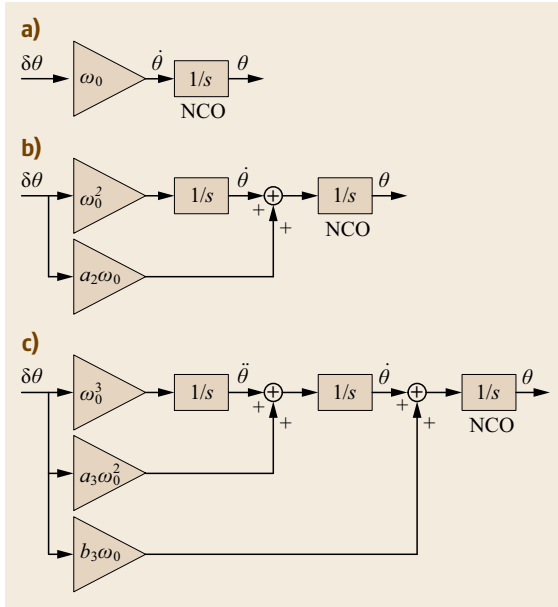


Fig. 14.12a–c Block diagrams of loop filters: (a) first order, (b) second order, and (c) third order (after [14.10])

filters

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{F}\mathbf{x}_k + \mathbf{L}\delta\theta_k, \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k + \mathbf{D}\delta\theta_k, \end{aligned} \quad (14.56)$$

where \mathbf{x} is the state vector, \mathbf{y} is the system output, $\delta\theta$ is the discriminator output, \mathbf{F} is the transition matrix, \mathbf{L} is the filter gain matrix, and \mathbf{C} and \mathbf{D} are the corresponding matrices to obtain the system output from the system state vector and input vector. All of these are dependent on the selected filter order and loop bandwidth.

In a similar way, the open-loop transfer function $F(s)$ in (14.42) can also be expressed in a state-space form for any order of loop filters.

Note that only the second equation of (14.56) is used within a real receiver implementation and links the discriminator output $\delta\theta$ to the NCO (code or carrier) phase and rate contained in \mathbf{x}_k .

Characteristics of the loop filters of three different orders are summarized in Table 14.2. In the implementation of $F(s)$, a subscript L like $(\cdot)_L$ is intentionally used to distinguish state-space matrices from continuous time symbols. Note that the state vector of the open loop is the rate information of the closed loop, therefore the order of \mathbf{x}_L is one order less than \mathbf{x} due to the NCO in the closed loop; for an example of third-order loop $\mathbf{x} = [\theta, \dot{\theta}, \ddot{\theta}]^T$ and $\mathbf{x}_L = [\dot{\theta}, \ddot{\theta}]^T$. In the general case of signal dynamics specifications the first-order loop filter is widely employed in the DLL with a Doppler aiding, whereas the second- and third-order

loop filters are eventually used in the FLL and PLL respectively.

Note that the tracking loops in the continuous time domain are unconditionally stable by nature but the corresponding discrete equivalents are a kind of sampled data system that is never unconditionally stable. This is due to the fact that high-gain loops in the discrete time domain always result in instability because of the inherent transportation lag, which is a major potential drawback of this kind of system. Moreover, the stability issue becomes serious when the normalized bandwidth (i.e., BT = the product of the loop bandwidth and the loop update time interval) in the discrete systems is insufficiently small compared to a certain threshold. Therefore, the stability issue is an important factor in any kind of digital signal tracking loop.

As a result, the key point in designing a tracking loop filter is to determine both the filter order and the filter noise bandwidth to accommodate user dynamic stress specification based on the noise statistics given by RF specifications of the receiver as well as the integration time [14.17].

14.4.6 NCO and Code/Carrier Generator

The signal parameter error sensed at the discriminator is filtered by the loop filter, which produces the signal parameter rate. This signal parameter rate is then applied to the NCO, which is a digital signal generator creating a synchronous, discrete-time, discrete-valued representation of a waveform. The NCO increases or decreases by the rate information as necessary to adjust the local replica signal generator phase with respect to the incoming signal phase. Therefore, mathematically, the NCO can be regarded as a simple integrator that gets the rate and then produces the phase. The obtained phase is then applied to the carrier generator to generate the appropriate carrier wave for the next step. A similar metric is also applied to the code NCO and generator.

Figure 14.13 illustrates a block diagram of the carrier NCO and its digital frequency synthesizer waveform, for example, high and low frequency trigonometric functions. One replica carrier cycle and one replica code cycle repeat each time when the corresponding NCO overflows. In particular for the carrier NCO, a map function that converts the amplitude of the NCO staircase phase output into the appropriate trigonometric function [14.10] is used even in modern digital real-time receivers due to computational efficiency. However, due to the evolution of digital processing techniques and the recent emergence of a software-defined-radio approach for GNSS receivers, a very high resolution NCO to generate a continuous

Table 14.2 Loop filter characteristics

	First order	Second order	Third order
Loop filter transfer function	$F(s) = \omega_0$	$F(s) = \frac{a_2\omega_0s + \omega_0^2}{s}$	$F(s) = \frac{b_3\omega_0s^2 + a_3\omega_0^2s + \omega_0^3}{s^2}$
Implementation of $F(s)$		$\mathbf{x}_L = [\dot{\theta}]$ $\mathbf{F}_L = [1]$ $\mathbf{L}_L = [T\omega_0^2]$ $\mathbf{C}_L = [1]$ $\mathbf{D}_L = [a_2\omega_0]$	$\mathbf{x}_L = [\dot{\theta}, \ddot{\theta}]^\top$ $\mathbf{F}_L = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}$ $\mathbf{L}_L = [Ta_3\omega_0^2 \quad T\omega_0^3]^\top$ $\mathbf{C}_L = [1 \quad 0]$ $\mathbf{D}_L = [b_3\omega_0]$
Closed-loop transfer function	$H(s) = \frac{\omega_0}{s + \omega_0}$	$H(s) = \frac{a_2\omega_0s + \omega_0^2}{s^2 + a_2\omega_0s + \omega_0^2}$	$H(s) = \frac{b_3\omega_0s^2 + a_3\omega_0^2s + \omega_0^3}{s^3 + b_3\omega_0s^2 + a_3\omega_0^2s + \omega_0^3}$
Implementation of $H(s)$	$\mathbf{x} = [\theta]$ $\mathbf{F} = [1]$ $\mathbf{L} = [T\omega_0]$ $\mathbf{C} = [1]$ $\mathbf{D} = [0]$	$\mathbf{x} = [\theta, \dot{\theta}]^\top$ $\mathbf{F} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}$ $\mathbf{L} = [Ta_2\omega_0 \quad T\omega_0^2]$ $\mathbf{C} = [1 \quad 0]$ $\mathbf{D} = [0]$	$\mathbf{x} = [\theta, \dot{\theta}, \ddot{\theta}]^\top$ $\mathbf{F} = \begin{bmatrix} 1 & T & 0 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix}$ $\mathbf{L} = [Tb_3\omega_0 \quad Ta_3\omega_0^2 \quad T\omega_0^3]^\top$ $\mathbf{C} = [1 \quad 0 \quad 0]$ $\mathbf{D} = [0]$
Loop noise bandwidth	$B_n = \frac{\omega_0}{4}$	$B_n = \frac{\omega_0}{1.89}$	$B_n = \frac{\omega_0}{1.2}$
Filter coefficients	1	$a_2 = \sqrt{2}$	$a_3 = 1.1, b_3 = 2.4$
Steady-state error	$\frac{dR/dt}{\omega_0}$	$\frac{d^2R/dt^2}{\omega_0^2}$	$\frac{d^3R/dt^3}{\omega_0^3}$
Characteristics	Sensitive to velocity stress	Sensitive to acceleration stress	Sensitive to jerk stress

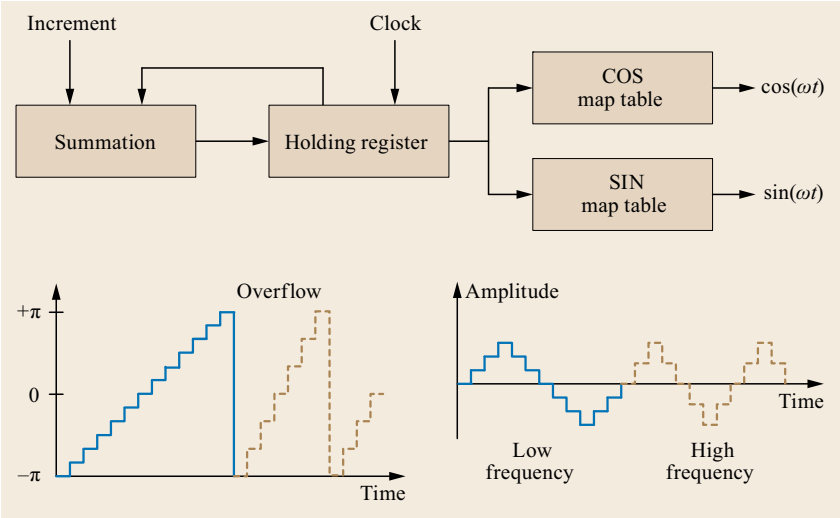


Fig. 14.13 Block diagram of digital frequency synthesizer based on NCO and its phase state and cos/sin map output

waveform based on a powerful processor is employed for the GNSS signal processing.

14.4.7 Aiding

As shown in Fig. 14.8 the use of aiding information on each tracking loop in a closed-loop operation enhances its performance. The aiding information is a rate of change of the corresponding signal parameter of interest; that is the velocity aiding is applied in order to enhance the tracking performance of the code delay or the carrier phase.

Carrier-Doppler Aided Code Tracking

There are two sources for measuring distance between the satellite and the user. The range information of the carrier phase is equivalent to the code delay at a much smaller noise level plus a bias. Therefore, the Doppler information from the carrier tracking loop can be used for estimating the code rate information after a proper scaling. In this case the loop bandwidth of the code tracking loop can be extremely narrowed. A reversal way to attempt to aid the carrier Doppler tracking by using the code rate would not be feasible due to the weak link property of the carrier tracking loop in terms of the receiver dynamics and thermal noise effects in comparison to the code tracking loop [14.10].

External Doppler-Aided Carrier Tracking

Alternatively, most of the signal dynamics can be captured by external aiding sensors – mostly inertial measurement units. Therefore, the velocity information from an external aiding sensor can be used for the carrier tracking with a quite narrow loop bandwidth. The velocity information obtained by an external sensor in the navigational frame is converted into the line-of-sight domain between the user and satellites, which is then applied to the carrier Doppler tracking in each channel. This external velocity aiding to the Doppler tracking will also help the code tracking in such a way as stated before.

14.4.8 Switching Rule

The signal processing starts with the acquisition. If a signal is present, which can be detected by testing the signal power, the signal processing enters the FLL, continuously checking the code lock. If the code and frequency are then locked the processing moves to the PLL, continuously checking the phase lock. If either phase lock or code lock is lost, then the processing goes back and proceeds again. This can be done by resetting (or reconfiguring) the weight factor of the PLL and FLL (Sect. 14.7.4).

However, for a short period immediately after the tracking loops start, they are especially sensitive due to the large uncertainty in the signal parameter estimates obtained from the acquisition (e.g., 180° for carrier phase, several hundreds Hz for Doppler and a half-chip for code). Therefore, the bandwidth of each tracking loop should be chosen to be sufficiently larger than the actual noise bandwidth of the input signal in the initial stage, in order for system robustness at the beginning. This is because of the fact that the signal tracking loop is a kind of multi-input multi-output (MIMO) system, which has three distinct input and outputs such as carrier phase, Doppler, and code delay, but each tracking loop is designed separately on the assumption of a single-input single-output (SISO) system with no error for other loops. This is only valid generally in steady state; for example a DLL is derived assuming PLL and FLL are perfectly working ($\delta\phi = 0$ and $\delta f_d = 0$), and PLL and/or FLL are designed in a similar way. Therefore, at the initial stage, each tracking loop has a certain amount of error that may be correlated with the error of other loops.

Furthermore, due to the nature of system dynamics, both the carrier tracking loop and the code tracking loop are sophisticated systems that will take a while to converge within the small error bound in the steady-state region after the transient time.

Figure 14.14 shows the independent step responses of the PLL, FLL and DLL with commonly used bandwidths in the proper region. Note that the settling time of the FLL is quite a bit longer than those of the PLL and DLL (e.g., the settling time of the FLL is larger than 4–5 s, whereas the PLL and DLL have a quite shorter settling time of less than 1 s). The FLL plays an important role in the tracking loop in the initial stage in terms of aiding capability to DLL and PLL. This is because the minimum required C/N_0 of FLL is low in comparison to the PLL and the Doppler-aided code delay estimate from the FLL is much less noisy than the unaided DLL; in another words, the PLL and DLL begin to work correctly after the FLL works properly.

These facts should be taken into account in an appropriate way such that a wide bandwidth is preferable to be accepted especially at the beginning of the loop filter operation and then switched to narrower ones after the transient time to retain the robustness. In particular, the bandwidth switching of PLL from wide to narrow should not be faster than the time when the FLL converges sufficiently to a steady state. If not, the phase error of $\pm 90^\circ$ will occur and the PLL may not work correctly. A third-order PLL assisted with a second-order FLL for the unaided carrier tracking, together with a quite narrowband first-order DLL aided by Doppler from the carrier tracking loop is used for civilian purpose GNSS receivers.

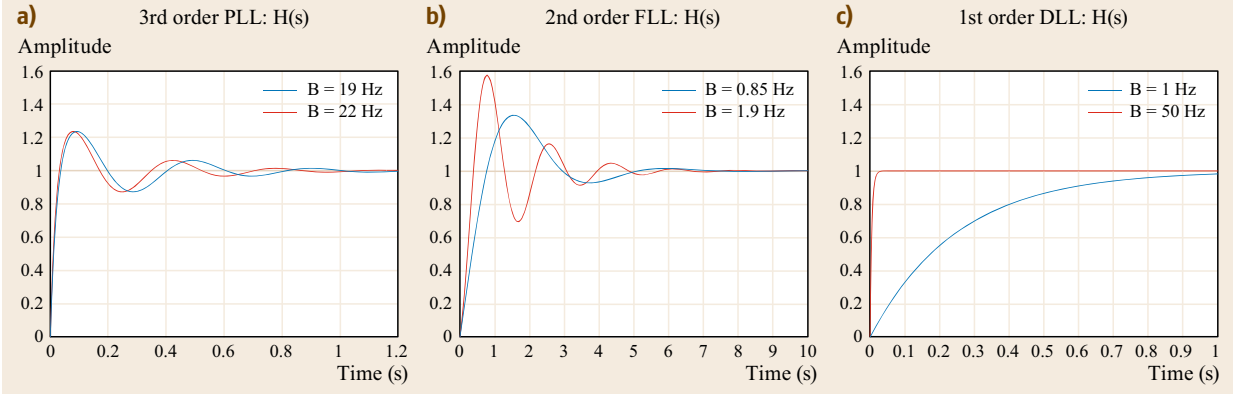


Fig. 14.14a–c Step response of tracking loops: (a) the third-order PLL, (b) the second-order FLL and (c) the first-order DLL (after [14.18])

14.4.9 BOC Tracking

Some of the upcoming GNSS signals have been designed to implement a BOC modulation that offers improved performance in terms of code tracking accuracy, robustness to interference, and interoperability, while offering a spectral separation from BPSK-like signals due to its split spectrum [14.19]. However, the main drawback of BOC-like signals is their multipeak autocorrelation function yielding multiple zero-crossing points in the code discriminator function that implies possible biased tracking (or false acquisition) if no special care is considered [14.3]. Therefore, the BOC ambiguity problem should be solved to force the receiver signal processor to lock onto the central peak. Several schemes for unambiguous tracking of BOC-like signals have been reported [14.3–6]. One of the most obvious ways for this is the bump-jump method. It uses two correlation channels like very early (VE) and very late (VL), in addition to the typical correlation channels early (E), prompt (P), and late (L) where the VE and VL correlators are separated from the P correlators by a code delay very close to half the inverse of the offset carrier frequency, that is one peak apart [14.4, 5]. In this method the tracker is controlled to jump in the appropriate direction when the VE and VL correlation values are consistently higher than the P correlation values. The number of correlation channels for VE and VL may increase with appropriate code delays as the order of BOC-like signals (e.g., MBOC or Galileo PRS) increases.

14.4.10 Tracking Performance

Under the assumption of no multipath, the dominant sources of tracking error in a tracking loop of a GNSS

receiver are composed of tracking jitter and dynamic stress error (induced by signal dynamics). The tracking loops may maintain their lock when the tracking errors are sufficiently well controlled to be less than thresholds corresponding to the rule-of-thumb design strategy of tracking loops; the $3\text{-}\sigma$ tracking jitter has to be reliably less than the pull-in range of the corresponding discriminator in order to guarantee more than 99.7% of lock condition. If this rule of thumb is violated, the satellite signal will most likely be lost [14.10]. The rule of thumb is written as

$$3\sigma_{\text{XLL}} = 3\sigma_{\text{j,XLL}} + \varepsilon_{\text{D,XLL}} \leq T_{\text{XLL}}, \quad (14.57)$$

where σ_{XLL} is the 1-sigma PLL, FLL or DLL tracking error, $\sigma_{\text{j,XLL}}$ is the 1-sigma phase, Doppler or code delay jitter from all sources except dynamic stress error, and $\varepsilon_{\text{D,XLL}}$ and T_{XLL} are the dynamic stress error and the rule-of-thumb threshold obtained from the pull-in range of discriminator in the corresponding tracking loop, respectively (see [14.10] for more details in mathematical expression of each error source).

Figure 14.15 illustrates the tracking error with respect to C/N_0 . There are six key factors in the tracking loop jitter plots (Table 14.3). The tracking loop should be designed to have low minimum C/N_0 to lock onto the signal, to have small noise jitter error, a narrow spreading width and a large design margin. Unfortunately, no method to simultaneously fulfill all these requirements exists.

Depending on the type of tracking loop, the main sources of tracking error as well as the tracking threshold are different. The tracking jitter of a PLL consists of the thermal noise error, oscillator's Allan variance error, vibration-induced oscillator error, ionospheric scintil-

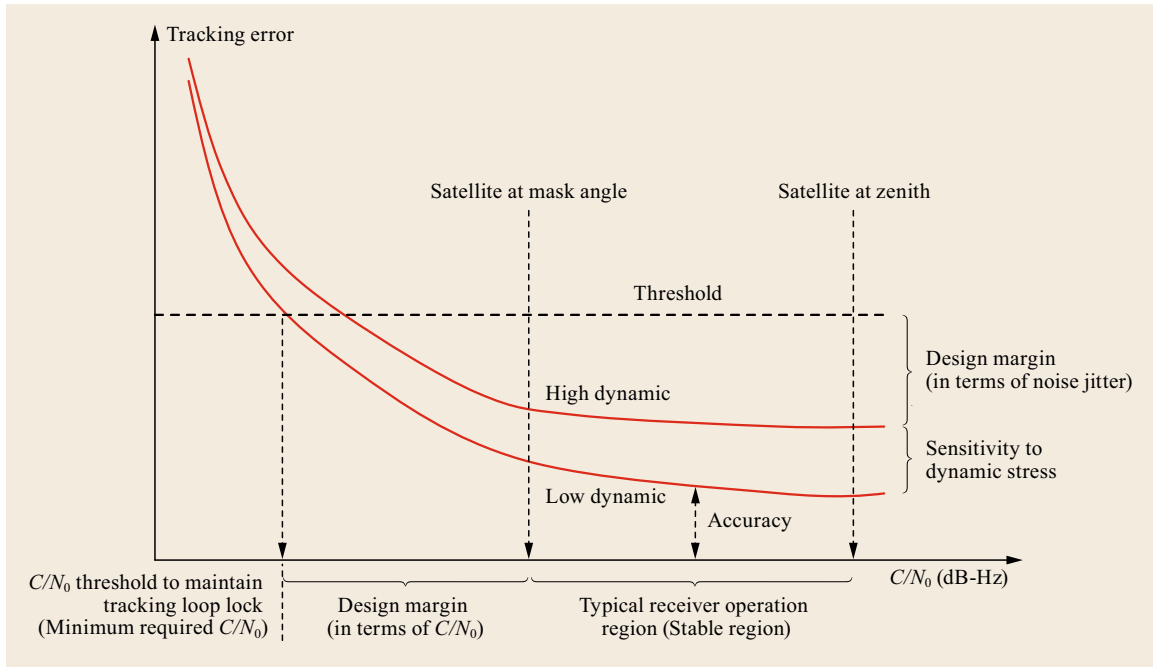


Fig. 14.15 An illustrative example of tracking loop error versus C/N_0

Table 14.3 Key factors in tracking loop jitter line

Key factors	Description
Minimal required C/N_0	The intersection point of noise jitter line with the threshold line; allowable C/N_0
Accuracy	A noise jitter at a certain C/N_0 in stable region
Dynamic stress sensitivity	Spreading width of noise jitter lines; sensitivity of tracking loop to dynamic stress; if the tracking loop is designed to be robust to signal dynamics the spreading width of noise jitter lines might be narrow for given dynamics
Design margin (in terms of noise jitter)	Threshold minus noise jitter at a certain C/N_0 ; signal tracking loop stability margin
Design margin (in terms of C/N_0)	Start point of stable region minus required minimal C/N_0
Stable region	Typical operating range of tracking loops; noise jitter line is aligned almost horizontally and produces a reasonable accuracy

lation error, and so on. The tracking jitter of FLL and DLL consists mainly thermal noise error.

The tracking threshold of a Costas PLL is normally 45° for a data channel and a pure PLL for a pilot channel has a 90° threshold.

The thermal noise error of a PLL is a function of the loop bandwidth and C/N_0 , while the FLL and DLL additionally require the contributions of the integration time T and the correlator spacing d , respectively. Both are also related to their tracking thresholds. Also, the thermal noise error of the tracking loops is slightly dependent on the discriminator algorithm employed and affected by the squaring loss that is dependent on the integration time T . The larger the C/N_0 is or the narrower the loop bandwidth is, the less thermal noise error is

effectively achieved. Furthermore, a longer integration time or a smaller correlator spacing results in a better tracking accuracy in terms of thermal noise.

It should be noted that, in particular, the thermal noise error of a DLL is affected by the shape of the power spectral density of the signal weighted with $(f - f_L)^2$. This weighted power spectral density is the Fourier transform of the second derivative $R''(\tau)$ of the autocorrelation function and its integral over frequency is called the *Gabor bandwidth*. This means BOC-like signals, occupying the same bandwidth as BPSK-like signals but having more signal power at the spectrum edge, have a larger Gabor bandwidth [14.20] and thus a sharper correlation peak. Therefore, the BOC tracking accuracy is better than that of BPSK-like signal.

The vibration-induced oscillator jitter of a PLL is caused by external vibration, which can be modeled by the oscillator's g-sensitivity, the vibration intensity and the frequency range of vibrations for different orders of loop filter as well as different carrier frequencies [14.21]. This can be a major problem for dynamic applications. Vibration isolators in mounting the oscillator should be taken into account to reduce its impact on PLL.

The oscillator's Allan variance jitter of a PLL is a natural phase noise resulting from frequency instability, which can be modeled by three (so-called) clock parameters (h_{-2} , h_{-1} , h_0) and the averaging time, depending on the oscillator grade for different orders of loop filter as well as different carrier frequencies [14.21, 22]. The oscillator's Allan variation jitter is proportional to the carrier frequency. The use of a high-grade oscillator is important in reducing the oscillator's Allan variation jitter.

The dynamic stress error is due to the relative motion between user and satellites, therefore it is a major problem for high-dynamic users, principally degrading the tracking performance. It is dependent on the signal dynamics, which is proportional to the carrier frequency, and the loop bandwidth for different orders of loop filter. The first-order loops are sensitive to velocity stress, while the second- and third-order loops

are sensitive to acceleration and jerk type stress respectively. It can be reduced if a lower carrier frequency and narrower loop bandwidths are used.

The ionospheric scintillation error of a PLL is due to ionization density irregularities, making signals change phase unexpectedly and varying rapidly in magnitude (see Sect. 6.3.3 of this Handbook). Ionospheric scintillation effects vary with an 11-year period of solar activity, mainly in low (tropical) and high latitude regions. The ionospheric scintillation effect can be modeled by an amplitude fade scintillation and phase variation scintillation. The strength of the amplitude scintillation is typically quantified by a metric called S_4 -index, which is the ratio of the standard deviation of the signal power to the mean signal power computed over a period of time [14.23]. In order to effectively track scintillating GNSS signals, it was demonstrated that a third-order PLL with $T \approx 10$ ms and $B \approx 10$ Hz provided good values for tracking of GPS L1 C/A code receivers in the presence of scintillations [14.24]. If the receiver application does not require carrier-phase lock, a FLL is preferable instead of a PLL since the FLL is more robust than a PLL during scintillation. Also a variable bandwidth PLL, designed to maintain lock during ionospheric scintillations, could be a good solution [14.25].

More detailed information on the tracking performance will be provided in Sect. 14.6.

14.5 Time Synchronization and Data Demodulation

By demodulating the navigation data message, the receiver is able to obtain all the necessary information for positioning (e.g., satellite ephemeris or atmospheric model parameters) and is also able to retrieve the transmission epoch of the satellite signal in an absolute sense.

The tracking process inside the receiver tries to steer $\delta\tau$ as defined in (14.10) towards zero. Thus the code phase \hat{v} of the replica signal, defined as

$$\hat{v}(k) = T_s k - \hat{\tau} \quad (14.58)$$

follows the code phase v of the received signal, defined as

$$v(k) = T_s k - \tau. \quad (14.59)$$

The code phase is per definition equal to the transmission epoch of the signal in the nominal satellite timescale modulo the code period. Thus by knowing the code phase, the transmission epoch of the GNSS signal

received at the epoch k is known modulo the code period.

To get the transmission epoch in absolute time, the number of code periods having already been broadcast by the satellite since a certain reference epoch (e.g., nominal week number 0 and second 0) has to be known. This number is sometimes called *code ambiguity*. The necessary information to achieve this is embedded in the navigation data message. For GPS it is composed of a week number and a so-called *Z-count* defining the time within the week in multiples of 6s. The latter term is called *time of week*. Other GNSSs use similar definitions. Overall those data fields define the transmission time of a certain bit boundary within the navigation message. Bit boundaries coincide with PRN code boundaries. After decoding the time information for this reference PRN code boundary, the channel keeps track of the received code periods and is always able to tell the transmission time in an absolute sense.

The process of resolving the code ambiguity can often be done without the message itself. The ambiguity

is rather long (typically at least $1\text{ ms} \approx 300\text{ km}$). Thus if the same satellite is tracked on a different frequency (e.g., already on L1), then the ambiguity on L5 for example can be easily derived from the first frequency. Other methods rely on consistency considerations making use of approximate user coordinates and satellite ephemeris data.

14.5.1 Bit/Symbol Synchronization

The binary units that are used to represent the navigation data message $d(t)$ are called bits or symbols. The term symbols is used in the case where the message employs a *forward error correction* scheme (like for the GPS civil navigation message (CNAV) on L2 or L5, or all Galileo messages). Otherwise the term bits is used (e.g., for the GPS NAV message on L1).

The number of primary code periods within one bit/symbol is an integer but can be larger than 1. Furthermore, a so-called secondary code sequence can be used, which alternates the sign of the primary code periods within one bit/symbol.

For the example of the GPS L5 data component, the secondary code is 1111001010. A 1 indicates to reverse the sign of the primary code, a 0 to keep the primary code sign. The length of the secondary code generally equals the bit/symbol duration. For the GPS L5 data component, the primary code has 10 230 chips, a duration of 1 ms and a data rate of 100 sps.

In case of the GPS C/A code signal on L1, the primary PRN code repeats 20 times within one data bit, which results in a trivial secondary code (= 00000000000000000000). All C/A code PRN sequences have the same sign within one bit.

A secondary code can also be present on pilot signal components.

The process of bit/symbol synchronization is sometimes also called secondary code synchronization. Once it has been achieved, the secondary code can be removed from the correlation values, which can then be added up coherently. This allows for increasing the coherent integration time from, for example, 1 ms to 10 ms, which is usually beneficial for the tracking process in terms of sensitivity and accuracy (the squaring loss decreases). Furthermore, the carrier phase can be retrieved unambiguously in the full range of $\pm 180^\circ$ as the sign of the secondary code is known.

A useful illustration of this process is shown in Fig. 14.16. Here the in-phase component of the data and pilot component of a rather clean Galileo E1 open service signal are plotted as a function of time. Phase lock has already been achieved and most signal power is in the in-phase component. The coherent integration time for data and pilot is here 4 ms corresponding to the primary code period.

At $t \approx 8.1\text{ s}$ the receiver achieved secondary code synchronization on the pilot channel. Before this epoch the secondary code (= 0011100000001010110110010) can be identified even visually on the pilot. After synchronization, the channel removes the secondary code from the correlation values, and corrects for the initially incorrect phase offset of 180° . The pilot correlation values change from being negative to the positive side.

There are several different algorithms possible to achieve secondary code synchronization. They can be grouped into the following categories:

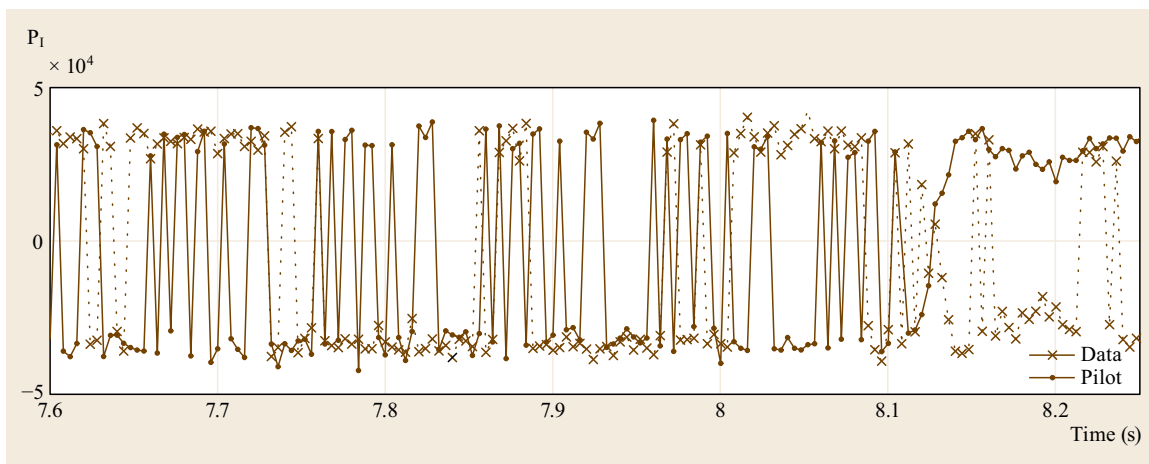


Fig. 14.16 Prompt correlator in-phase values while symbol synchronization is achieved on the pilot for an E1 Galileo open service satellite signal

- Histogram method (GPS C/A or Russian Global Navigation Satellite System (GLONASS) only)
- Postcorrelation search
- Dedicated tracking verification channels.

The histogram method works for all signals with a trivial secondary code. The algorithm detects bit transitions between primary code boundaries using for example the bit-transition-sensitive discriminator explained later in (14.61). Assuming a certain run time R of the algorithm (e.g., for 1 s runtime, $R = 1000$ in the case of GPS C/A), a histogram is set up with the number of bins L equal to the number of primary codes within one data bit (e.g., $L = 20$ in the case of the GPS C/A signal). Each bin n counts the events of $\text{dot}_k \bmod L = n < 0$ using the definition (14.61). Here $k \bmod L$ denotes k modulo L . At the true bit transition bin n_t , there will be on average $R/(2L)$ counts. The other bins generally contain zero counts for high signal power. It is relatively easy to find appropriate thresholds to identify the correct bin transition epoch.

For the postcorrelation method, the tracking channel produces prompt I/Q correlation values based on a coherent integration time equal to the primary code period. Let now s_k denote the secondary code sequence. Assuming again a certain runtime R of the algorithm, the following test metric can be used to identify the true bit/symbol transition

$$\hat{n}_t = \arg \max_{n_t} \sum_{k=1}^{R/L} \left| \sum_{l=1}^L (I_{P,Lk+l-n_t} + jQ_{P,Lk+l-n_t}) s_l \right|^2. \quad (14.60)$$

The postcorrelation method can be extended to estimate also residual Doppler frequency errors, which are typically present at this point of signal tracking as the Doppler estimate from the acquisition engine is rather crude [14.26].

The disadvantage of the postcorrelation and the histogram method is that the underlying tracking channel operates with the short coherent integration time determined by the primary code period. It is often only 1 ms, which makes it difficult to track the signal in for example moderate indoor conditions or under canopy. A solution to circumvent this problem can be realized if an abundance of tracking channels is available (which is not a problem for today's application specific integrated circuit (ASIC) or CPU technology). Then not only one tracking channel is started per acquired PRN code, but L channels are started and each channel assumes a certain bit/symbol transition epoch (the number of hypothesized equals to L). Each channel employs a much longer integration time (e.g., 20 ms). The chan-

nel with the correct hypothesis will finally turn out the highest estimated signal power and after that the other channels are stopped.

14.5.2 Data Bit/Symbol Demodulation

Once bit/symbol synchronization has been achieved, the navigation data message can be stripped off from the received signal. This is rather straightforward and three major scenarios can be considered.

The most direct way occurs if a pilot signal is present and phase lock has been achieved (e.g., Fig. 14.16). Then the in-phase prompt correlator values $I_{P,k}$ of the data-bearing component correspond to the bits or symbols (the index k is used to enumerate the sequence of values). Usually a positive in-phase correlator value is attributed to a bit value of 0 and a negative correlator value to a bit value of 1. The case of a zero in-phase correlator value is very unlikely to occur and can be handled either way.

The second scenario is the case of no pilot signal being present but phase lock having been achieved using a Costas PLL. The bits/symbols can be again retrieved from the sign of the in-phase prompt correlator values $I_{P,k}$, but one doesn't know if the navigation message is inverted or not. A possible inversion is resolved later in the decoding process.

Finally, the third scenario corresponds to the case where frequency lock is achieved but no phase lock. Then the sign of the in-phase prompt correlation values cannot be used directly for demodulation. Instead one computes phase transitions between two epochs like the dot term of the FLL discriminator,

$$\text{dot}_k = I_k I_{k+1} + Q_k Q_{k+1}. \quad (14.61)$$

If the dot term is positive, the assumed bit/symbol sequence keeps the same sign, if it is negative the sign changes. Like for the second scenario, one does not know if the resulting bit/symbol sequence is inverted or not. Furthermore an erroneous estimate of one transition inverts the whole resulting bit/symbol stream from that epoch on. Frequency tracking can be applied for lower received signal powers compared to phase tracking. But still the frequency tracking errors need to be reasonable, in order not to mix up residual frequency contributions with bit/symbol transitions.

The bit/symbol error rate can be evaluated analytically if one assumes perfect phase or frequency tracking. In that case, bit/symbol errors occur due to the fact that $I_{P,k}$ and $Q_{P,k}$ are nonzero Gaussian random variables. If the noise exceeds the signal amplitude, bit/symbol errors occur. For the first and second sce-

nario, the bit/symbol error rate (BER) is given by

$$\text{BER}_{\text{PLL}} = \frac{1}{2} \text{erfc} \sqrt{TC/N_0} \quad (14.62)$$

with $\text{erfc}(x) = 1 - \text{erf}(x)$ denoting the complementary error function. For the third scenario the bit/symbol transition error rate (TER) is given by

$$\text{TER}_{\text{FLL}} = \frac{1}{2} \exp(-2TC/N_0). \quad (14.63)$$

Generally, the thermal noise contribution to the bit/symbol error rate is rather small and errors are more likely to occur due to tracking errors, occasional blocking of the line-of-sight or signal fading effects. Those contributions can, however, not be tackled in a simple analytic way [14.27].

14.5.3 Frame Synchronization

The various navigation messages have their own structure, but the data is generally organized in blocks called frames, pages, subframes or similar. One core task of the decoder is to identify the beginning of such a block. This is facilitated by a so-called preamble. The preamble is a well-defined bit/symbol sequence broadcast at the beginning of the block. A few exemplary preambles are shown in Table 14.4.

For most navigation messages (e.g., GPS C/A NAV, GLONASS, Galileo, Beidou), the preamble can be found in the bit/symbol stream directly after demodulation. That is, no forward error correction scheme as described in Sect. 14.5.4 is employed for it. Prominent exceptions are the satellite-based augmentation system (SBAS) and the GPS L2/L5 CNAV messages, where the forward error correction scheme also includes the preamble. In that case, the Viterbi decoder has to be applied first, before searching the preamble.

If the absolute sign of the navigation message is not determined by the demodulation, two hypotheses for the preamble have to be tested. The normal and the inverted preamble have to be searched.

Once the preamble has been identified, a predefined number of bits/symbols representing the block of the navigation data message is passed to the decoder. If the

inverted preamble has been found, the block has also to be inverted. If the decoder is able to validate this block of data, by for example employing parity checks, the channel achieves *frame synchronization*. Usually a few trials are necessary, as the preamble is rather short and it is likely that also part of the block data is mistaken for it.

14.5.4 Bit Error Correction

To improve the resilience of the receiver against bit/symbol errors caused by distortions in the signal propagation channel, various measures have been employed by the GNSS signal designers. Those techniques reduce the bit error rate (error correction) and provide means to verify that bits have been received correctly (error detection):

- Parity bits
- Forward error correction
- Interleaving.

For example, the smallest data block of the GPS C/A NAV message (called a word) consists of 30 bit and six of them are parity bits. The underlying algorithm to compute parity is given in [14.28]. For more recently designed navigation messages, parity is often computed using cyclic redundancy checks (CRC), which are well suited to detect burst errors (contiguous sequences of erroneous data) [14.29, 30].

To allow also the correction of erroneously transmitted symbols, so-called *forward error correction* (FEC) schemes are employed.

Typically in GNSS a FEC 1/2 with a constraint length of 7 is used. The broadcast navigation data bits are convoluted with two different bit sequences (often called polynomials in this context) of length 7. The output of those two convolutions is timely multiplexed resulting in the symbol stream. The symbol stream has twice the data rate of the bit stream. The encoding can be performed in a continuous way over the whole navigation message (e.g., GPS CNAV, SBAS) or over selected block of the message (e.g., for Galileo messages). The receiver employs a so-called Viterbi decoder to decode the symbols and to retrieve the data bits from it. The decoder is not only able to identify bit errors but is also able to correct them to some extent. The Viterbi decoder causes a delay in the data stream, which has to be accounted for when the time of week is retrieved from the message.

As the Viterbi algorithm has generally a better performance if symbol errors are equally distributed in the data stream but burst errors are more likely to occur during the transmission, so-called interleaving is

Table 14.4 Navigation message preambles for GNSS signals near the L1 frequency

Message	Preamble
GPS NAV L1	1000101100
GLONASS	111110001101110101000010010110
Galileo INAV E1	0101100000
Beidou D1	11100010010

employed. The symbols within one block of data are reshuffled in a defined way, such that burst errors are spread out. The receiver has to rearrange the symbols before passing them to the Viterbi decoder.

The Galileo INAV messages transmits identical data on E1 and E5b but in a timely different order. So-called even and odd pages alternate on the two signals. This allows a dual-frequency receiver to decode the navigation message in a shorter time frame and also increases the robustness against burst errors.

A GNSS navigation message changes its content typically only every few hours, when new ephemeris data is uploaded from the ground segment. It is therefore possible to increase the decoding sensitivity of a receiver by stacking multiple instances of the received data stream. Furthermore, the message can be

relatively well predicted ($\approx 99\%$) once it has been received completely. This generally increases the robustness for carrier-phase tracking as four-quadrant phase discriminators can be employed instead of Costas discriminators.

14.5.5 Data Extraction

Once the message has been received and all parity checks have passed successfully the data content can be read. It consists of a sequence of bit-fields and each field typically represents a signed or unsigned integer number. Those integers are converted to floating point numbers using scale factors and offsets. The precise algorithms can be found in the interface control documents (ICDs).

14.6 GNSS Measurements

The primary measurements of a tracking channel are:

- The estimated code pseudorange between satellite and receiver
- The estimated Doppler of the received signal
- The estimated carrier phase (or carrier pseudorange) of the received signal
- The estimated amplitude (or power) of the received signal.

These four values are generated for each tracked GNSS signal by the receiver. If the receiver tracks multiple signals or services of one satellite, then independent values are generated for each signal. The output rate is application dependent and usually ranges from 1–20 Hz. The measurements are taken at exactly the same epoch for all tracking channels. If a navigation signal consists of a data and pilot component, then it is common practice to output one single measurement set incorporating contributions from both components, as most of the important error sources (multipath, transient errors or hardware delays) are highly correlated between these components anyway. Only the carrier pseudorange might be based on the pilot component only, as this is free of the 180° ambiguity.

Usually the measurement epoch is aligned to GNSS time (most commonly to the GPS time). This ensures that data from different receivers can be properly aligned to each other, which could otherwise become tricky if each receiver would rely on its own clock. The measurements of two different receivers would tend to shift with respect to each other.

The measurements reflect the true receiver position, velocity and the receiver clock behavior. Furthermore,

several atmospheric effects, receiver-satellite hardware components and signal processing artifacts influence the measurements. The relation between those unknowns and the measurements are called *observation equations*. They will be discussed in the following sections from the signal processing point of view. A more generic discussion can be found in Chap. 19.

14.6.1 Code Pseudorange

Loosely speaking, the *code delay* of the replica signal $\hat{\tau}$ follows the true but unknown delay τ of the received signal. Looking in more detail into the correlation process defined in (14.18) one realizes that the actual argument of the spreading sequence is the *code phase* of the replica signal given by $\hat{v}(k) = T_s k - \hat{\tau}$. The code delay multiplied with the speed of light is the *code pseudorange*. It is also referred to as *code* or *pseudorange* for simplicity.

It is common practice that the receiver uses the code phase $\hat{v}(k)$ as defined in Sect. 14.5 as a working variable for PRN code generation and not the code delay. This is mostly related to the fact that the code phase is directly related to the generated spreading code, which saves gates in an ASIC and/or CPU power. Working with separate variables for $T_s k$ and $\hat{\tau}$ would also require a high number of bits to represent the covered time spans with the required resolution of less than 1 mm. Overall the receiver maintains the code phase $\hat{v}(k)$ of (14.58) in one register of the NCO. This code phase is split into a fractional part being smaller than one code period, and the number of code periods as described in Sect. 14.5. This combined

code phase equals the nominal sent time of the signal.

If the considered GNSS signal has a data and a pilot component of a different primary PRN code length, it might be reasonable to maintain separate code phase entries in the NCO. If combined data and pilot processing is considered as described in Sect. 14.7.3 then the code and pilot NCO values are kept synchronously. That is, the data/pilot code phases modulo the shorter PRN period are identical.

With these considerations the code pseudorange $p(k)$ expressed in meters can be obtained via

$$p(k) = c\hat{\tau}(k) = c[t_{\text{rec}} - \hat{v}(k)] \quad (14.64)$$

and it is the difference of the nominal receiver time t_{rec} minus the estimated sent time $\hat{v}(k)$ multiplied with the speed of light c .

The nominal receiver time is the raw receiver time. It is realized by simply counting the samples produced by the ADC, thus $t_{\text{rec}} = T_s k$. The ADC sampling rate (as well as the local oscillators for downconversion) is derived from the receiver oscillator. The oscillator may be of different quality ranging from a simple quartz to an atomic frequency standard. The nominal receiver time has an arbitrary offset to the true time, called receiver clock error, and this offset changes with time according to the oscillator drift.

If the measurement epoch (being aligned to e.g., a full GPS second) falls between two sampling epochs, then linear interpolation is sufficient to obtain both time values. This interpolation is easily done in the receiver firmware based on the code phase and code rate values of an adjacent sampling epoch.

The pseudorange definition (14.64) is sometimes called *absolute pseudorange* as it can be done for each tracking channel individually. Sometimes it is more convenient to introduce another definition – *relative pseudoranges* – which are based on choosing a reference channel denoted with an subscript 0. Then the relative pseudorange is defined as

$$p_r(k) = c(\hat{\tau}(k) - \hat{\tau}_0(k)) = c[\hat{v}_0(k) - \hat{v}(k)] \quad (14.65)$$

It can be computed without decoding the navigation message and resolving the code ambiguity, as the code phase difference between two tracking channels is readily available inside the receiver. The relative pseudorange of the reference channel is zero. The relative pseudorange does not contain any receiver clock error, thus timing information cannot be extracted from it. The relative pseudorange is rarely used and only if utmost simplicity is sought for receiver design.

The absolute pseudorange is usually modeled as

$$p = \rho - c(dt_r - dt^s) + I + T + c(d_r + d^s) + p_T + \epsilon_{\text{mp}} + \epsilon_n \quad (14.66)$$

Here ρ denotes the geometric distance between the receive antenna at the reception epoch and the transmit antenna at the transmission epoch. dt^s and dt_r are the satellite and receiver clock offsets with respect to a common system timescale, I and T represent ionospheric and tropospheric path delays, d_r and d^s denote satellite- and receiver-specific group delays, and the remaining terms describe various forms of measurement errors. These include loop transient errors (p_T), multipath errors (ϵ_{mp}) and tracking noise (ϵ_n). All terms generally depend on the time or epoch k .

The observation equation introduced in Chap. 19, namely (19.6), describes the same model as (14.66) but from the positioning point of view. In particular (19.6) uses the indices r , s and j to denote different receivers, satellites or signals, whereas (14.66) omits indices as it is for one tracking channel and thus for one receiver, one satellite and one signal. The measurement error e in (19.6) equals the sum of $p_T + \epsilon_{\text{mp}} + \epsilon_n$ in (14.66). Furthermore, the relativistic correction δt^{rel} of (19.6) is merged with dt^s in (14.66) and phase center variations ξ of (19.6) are neglected here.

The receiver clock error $dt_r(k)$ is the deviation of the nominal receiver time $t_{\text{rec}} = T_s k$ from the true GNSS timescale $t_t(k)$,

$$dt_r(k) = T_s k - t_t(k) \quad (14.67)$$

The receiver clock may assume numerically very large values, because the nominal receiver time does in general not reflect the absolute start time of the measurement. Instead, the start of the measurement expressed in the nominal receiver time is zero, since $k = 0$. To avoid numerical problems when producing the code pseudorange, the GNSS receiver introduces an artificial and additional clock error $dt_{r,a}$, which is applied before outputting the code (and also the carrier) pseudorange according to

$$p(k) \rightarrow p(k) - c dt_{r,a}(k) \quad (14.68)$$

The artificial clock $dt_{r,a}$ error usually assumes integer multiples of 1 ms and may change over time to keep the combined receiver clock error $dt_{r,a}(k) + dt_r(k)$ close to zero (and the pseudoranges near 25 000 km). Every time the artificial clock error changes its value one speaks of a *receiver clock jump* or 1 ms jump. The artificial clock error might also be adjusted continuously and steered towards zero, but discrete jumps are easier to detect

in the navigation software and also allows real physical modeling of the receiver clock error. Both cases can be seen as some kind of *software* clock steering. Many receivers are equipped with oscillators that can be tuned via a configurable input voltage. If this approach is used, one speaks of *hardware* clock steering. Hardware clock steering needs also to be modeled in the navigation process.

The satellite clock error dt^s in (14.66) measures the deviation of the transmission time $\nu(k)$ in the nominal satellite timescale to the true GNSS timescale,

$$dt^s(k) = \nu(k) - t_t(k). \quad (14.69)$$

It is broadcast by the navigation message and kept low (usually much below 1 ms) by the control segment.

The ionospheric delay is denoted as $I(k)$ and the tropospheric delay by $T(k)$. Both are positive and more details can be found in Chap. 19.

The symbol d_r in (14.66) denotes the receiver hardware delay. It is composed of contributions from the antenna, cables, amplifiers, mixers and filters. The receiver hardware delay is temperature dependent and thus generally time dependent. For signals with an identical modulation scheme, for example for GPS L1 C/A code signals from all satellites, it is usually considered to be identical. There are small differences depending on the modulation scheme (e.g., between BOC and BPSK signals) if the same center frequency is used. Larger differences occur for different center frequencies (e.g., for the GLONASS G1 signals), because group delay variations of the RF filters cannot be neglected for surface acoustic waves (SAWs) or ceramic filters, which are commonly used in GNSS receivers.

The satellite hardware delay is denoted as d^s and similar considerations as for the receiver hardware delay apply.

The symbol ϵ_{mp} is used to denote the ranging error caused by the fact that in general not only the line-of-sight signal is received but also reflections from nearby objects. This so-called *multipath* is for many applications the most difficult error to get control of and is discussed also in Chap. 15. Every single object in the surroundings of the receiving antenna reradiates reflections from the satellite signal. Although the single contributions might be small, their accumulation can cause a generally increased noise budget, which can easily achieve a few meters standard deviation. There might also be cases of a large surface acting as a specular reflector causing a single dominant multipath reflection. This case can be handled in a relatively simple theoretical way by means of so-called *multipath envelopes*. Assuming a single multipath reflection of a relative amplitude α and an increased path length τ_m

(both with respect to the line-of-sight signal), the corresponding ranging error can be bounded by writing

$$t_{mp;-}(\alpha, \tau_m) \leq \epsilon_{mp} \leq t_{mp;+}(\alpha, \tau_m). \quad (14.70)$$

The bounds $t_{mp;\pm}(\alpha, \tau_m)$ are obtained by solving the following equation for t

$$\text{real}[D_c(t) \pm \alpha D_c(t + \tau_m)] = 0. \quad (14.71)$$

Here $D_c(t)$ is the coherent part of the used code discriminator. The coherent part is the complex valued discriminator computed for example by subtracting the complex valued late correlator from the early correlator. The ranging error bounds $t_{mp;\pm}$ equal the resulting value of t and depend on τ_m and α . For $\tau_m = 0$, the multipath error vanishes. The two signs considered are either for constructive or destructive interference of the line-of-sight signal with the multipath reflection and represent the maximum magnitudes of the multipath error. The two solutions are denoted as $t_{mp;\pm}(\alpha, \tau_m)$ and are chosen so that $t_{mp;+} \geq t_{mp;-}$ (which can always be achieved).

An example for a CBOC signal with a narrow correlator is shown in Fig. 14.17. By choosing a narrow correlator spacing d , the maximum multipath error can be reduced. By combining two early/late pairs into one code discriminator, the medium and far range multipath can be virtually eliminated (at least for BPSK signals). There are many other ways to optimally select the code discriminator, which clearly goes beyond the scope of this chapter. For further reading see for example [14.31]. Near-range multipath with delays of maximally several meters still remains a big hurdle in navigation signal processing.

The symbol ϵ_n in (14.66) denotes the noise in the measurement. The noise comprises thermal noise

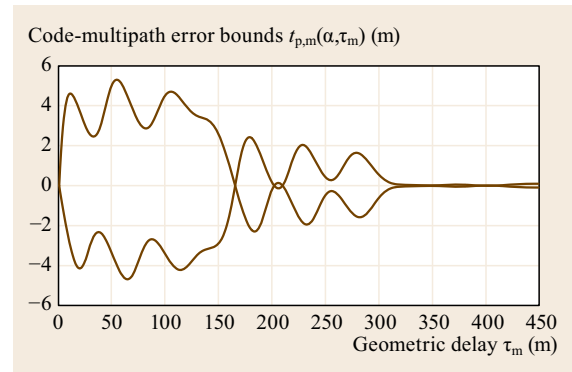


Fig. 14.17 Code multipath error envelope for a 12 MHz Galileo E1B signal, an early/late correlator spacing of $d = 0.05$ chip and a $\alpha = -6$ dB multipath signal

contributions mostly from the first low-noise amplifier (LNA) after the passive antenna element and quantization noise of the ADC. If the RF noise can be considered as white noise, which is described by a noise density N_0 , then ϵ_n is a zero mean Gaussian random variable whose variance depends on signal type, discriminator type and loop parameters. For the example of an idealized infinite bandwidth BPSK signal, and an early-power minus late-power code discriminator (14.55) with an early-late spacing of d chips, the variance σ_n^2 of ϵ_n expressed in squared meters is according to [14.16]

$$\sigma_n^2 = T_C^2 \frac{B_{\text{DLL}} d}{2C/N_0} \left[1 + \frac{2}{(2-d)TC/N_0} \right]. \quad (14.72)$$

The symbol T_C denotes the chip duration in meters.

A lower DLL bandwidth B_{DLL} and a high signal power C reduce the noise. The term in the brackets is called squaring loss and is relevant for low received signal power, which typically occurs if the receiver is not operated in open sky conditions. By increasing the coherent integration time T , the squaring loss can be reduced. This comes at the cost of increased receiver complexity, especially if T is larger than the data bit/symbol duration. Commercial receivers often employ integration times of a few milliseconds.

Though a reduction of the correlator spacing d also diminishes the noise, this reduction levels off for finite bandwidth signals and small values of d . In fact (14.72) is only an approximation and for finite bandwidth signals a rather complex theory exists to compute the variance. It relates the power spectral density of the signal to the power spectral density of the received noise in dependence on the used code discriminator [14.7, 32]. This theory is then also able to give closed expressions for BOC signals or other modulation schemes.

The spectral characteristics of the measurement noise ϵ_n relates to the DLL bandwidth B_{DLL} and the measurement rate. In case the product of both is much larger than 1, then ϵ_n can be assumed to be white. If near or lower than 1, ϵ_n is *colored noise* and ϵ_n reflects the filter characteristics of the DLL. In that case, the noise power is contained between 0 Hz and B_{DLL} but (14.72) is still valid. As a consequence, receivers with a high measurement rate need to employ high tracking loop bandwidths, as they would otherwise (i.e., with low loop bandwidths) generate timely correlated measurements.

For the already considered Galileo signal E1B, the code noise is shown in Fig. 14.18. At a typical signal power value of 45 dB-Hz it is two decimeters and for many applications negligible compared to the multipath error.

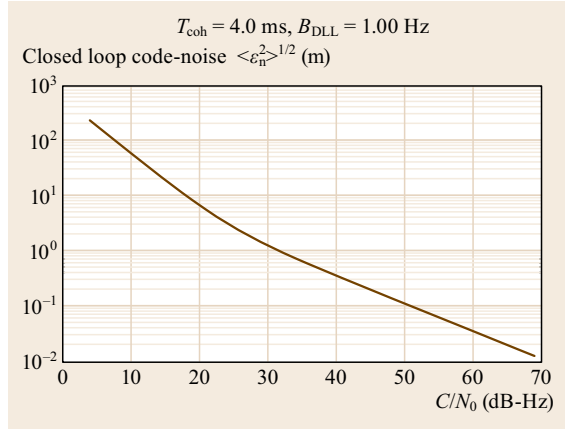


Fig. 14.18 Code noise for a 12 MHz Galileo E1B signal and an early/late correlator spacing of 0.05 chip

The *transient* error p_T is the opponent of the thermal noise error ϵ_n as both cannot be minimized at the same time. The transient error reflects the inability of the tracking loop to perfectly follow the signal dynamics (user motion) and the tracking loop lags behind the true signal dynamics. Transient errors are, due to the nature of the user motion in general, timely correlated but shall not be mistaken for colored measurement noise, as they are independent of the measurement rate.

For a first-order DLL, the response of the tracking loop to a user range jump of 1 m is

$$p_T = e^{-4B_{\text{DLL}}t} \quad (14.73)$$

and shown in Fig. 14.14. A lower loop bandwidth B_{DLL} causes transient errors to persist for a longer time. Transient errors exist also for higher-order loop filters and receiver design is always a trade-off between thermal noise and transient errors. B_{DLL} is chosen to be application dependent. For example, a second-order DLL with a bandwidth of 0.2 Hz may be used in a high-sensitivity mass market receiver and is able to track signals as low as ≈ 15 dB-Hz. In that case transient errors are experienced by the user, even for typical vehicular accelerations.

14.6.2 Carrier Phase

The locally generated replica signal as defined in (14.5) has an instantaneous carrier phase $\phi_{\text{NCO}}(k)$ defined as the argument of the complex exponential. It is one variable of the NCO and related to the definitions of (14.5) by

$$\hat{\phi}_{\text{NCO}}(k) = 2\pi \left(f_{\text{IF}} + \hat{f}_d \right) T_s k + \hat{\phi}(k). \quad (14.74)$$

Similar to the code pseudorange, the NCO does not work with $\hat{\phi}(k)$, but rather with $\hat{\phi}_{\text{NCO}}(k)$, which is the direct argument of the sin/cos functions. Otherwise the multiplication of $(f_{\text{IF}} + \hat{f}_{\text{d}})T_s k$ may cause numerical errors or may consume unnecessary area on the GNSS chip (or unnecessary instructions in the processor) to realize the multiplication, especially if large time spans are considered.

If phase tracking is employed and phase lock is achieved, this instantaneous carrier phase follows the received carrier phase, as the PLL steers

$$\hat{\phi}_{\text{NCO}}(k) - 2\pi(f_{\text{IF}} + f_{\text{d}})T_s k + \phi(k)$$

towards zero.

Referring back to the discussion of the code pseudorange in the previous section, the absolute *carrier pseudorange* $\varphi(k)$ expressed in meters is defined as

$$\varphi(k) = -\lambda \left(\frac{\hat{\phi}_{\text{NCO}}(k)}{2\pi} - f_{\text{IF}} T_s k \right). \quad (14.75)$$

The carrier pseudorange, which is most commonly called *carrier phase* for simplicity, is the difference between the NCO carrier phase minus the nominal increase (provided that $f_{\text{IF}} \neq 0$). The latter term can also be seen as the nominal receiver time scaled with f_{IF} and the first term is the transmitted carrier phase from the satellite. Carrier-phase measurements are taken simultaneously with code measurements eventually employing interpolation techniques if the measurement epoch is between two samples.

If the GNSS signal has a data and pilot component two options exist to generate the carrier pseudorange: carrier-phase measurements can be taken separately for data and pilot or a combined measurement can be generated, as further detailed in Sect. 14.7.3. As the nominal phase difference between data and pilot signal is known (usually it is 0° or 90°), and both components are affected by the same propagation delays, there is no information loss if only a combined measurement is generated. An important exception from this rule can occur, when utmost accuracy is sought in determining receiver-satellite hardware delays (as it is done in the control segment and networks of reference GNSS receivers). If the satellite payload combines data and pilot component using analog circuits then, a residual phase bias between data and pilot may exist due to inevitable tolerances in the circuits of a few degrees. This bias needs then to be estimated, for carrier-phase-based applications.

A possibly applied artificial receiver clock error (14.68) must also be accounted for in the carrier-

phase measurements and we write

$$\varphi(k) \rightarrow \varphi(k) - c dt_{\text{r,a}}(k). \quad (14.76)$$

To keep the numerical range of the carrier-phase measurements limited, one may start counting the carrier phase from a reference epoch $t_{0,c}$ on. This results in

$$\varphi(k) = -\lambda \left(\frac{\hat{\phi}_{\text{NCO}}(k)}{2\pi} - f_{\text{IF}}(T_s k - t_{0,c}) \right). \quad (14.77)$$

The epoch $t_{0,c}$ can be chosen quite freely, and shall be kept constant as long as the carrier phase is within the valid numerical range. Usually one tries to keep the carrier-phase values near to the pseudorange values. When choosing $t_{0,c}$, one should make sure that

$$e^{2\pi j f_{\text{IF}} t_{0,c}} = 1 \quad (14.78)$$

to ensure that the integer character of the carrier-phase ambiguities is not changed, when readjusting $t_{0,c}$.

The carrier phase is modeled as

$$\varphi = \lambda N + \rho - c(dt_{\text{r}} - dt^{\text{s}}) - I + T + c(\delta_{\text{r}} + \delta^{\text{s}}) + p_{\text{T}} + \epsilon_{\text{mp}} + \epsilon_{\text{n}}. \quad (14.79)$$

This model is similar to (14.66), with the exception that the ionospheric delay I changes sign and the carrier-phase ambiguity N has been added.

Comparing the model (14.79) to (19.9) of Chap. 19, we realize slight differences, which are due to the different focus of the respective chapters. Similar to the code pseudorange model (19.6), also (19.9) includes relativistic effects δt^{rel} and phase center corrections ζ . Both have been neglected in (14.79) as well as the phase wind-up correction ω . The measurement error ϵ of (19.9) equals the sum of $p_{\text{T}} + \epsilon_{\text{mp}} + \epsilon_{\text{n}}$ in (14.79).

The ambiguity term N can be seen as an integer (remember, satellite and receiver hardware delays δ_{r} , δ^{s} are modeled separately) and reflects the fact that $\hat{\phi}_{\text{NCO}}$ is initialized with an arbitrary integer part, when the tracking process starts. Furthermore the choice of $t_{0,c}$ is also in some sense arbitrary. Thus a 360° , or integer-cycle, ambiguity exists.

In the case where a data-only signal is tracked with a Costas PLL, the phase discriminator is insensitive to 180° jumps and the resulting half-cycle (180°) ambiguity can only be resolved by analyzing the preamble (Sect. 14.5.3) or other parts of the broadcast navigation message. Its correct resolution is essential for many navigation algorithms and receiver manufacturers put large efforts into this issue.

Unfortunately rather small effects like unexpected high user dynamics, short signal blocking, high multipath or high noise may disturb the PLL and phase lock can ultimately be lost. Even if this occurs only for a fraction of a second and phase lock is established again, the integer value of N may have changed, as the PLL only readjusts the fractional part but not the integer part. A *cycle slip* occurred. Its detection, and its eventual correction, is difficult, especially in a dynamic scenario. Cycle slips still are the main hurdle to be taken to fully exploit the otherwise very precise carrier-phase measurements.

The geometric range ρ and the receiver dt_r and satellite clock errors dt^s are the same as for the code pseudorange.

The receiver hardware delays δ_r and the satellite δ^s are similar to the code counterparts but may differ slightly depending on hardware components.

The transient error p_T reflects the inability of the PLL to instantaneously follow the user dynamics or receiver oscillator jitter and can be modeled similar to (14.73). Transient errors should be kept smaller than a fraction of a cycle by choosing a sufficiently high loop bandwidth to avoid cycle slips. Due to the short carrier wavelength, this limits PLL bandwidths to be greater than 5–10 Hz, which is much larger than for the DLL.

Multipath errors $\epsilon_{mp'}$ are omnipresent and typically much larger than thermal noise $\epsilon_{n'}$ errors. They are limited by a quarter of a wavelength $|\epsilon_{mp'}| \leq \lambda/4$ as detailed in Chap. 15. The maximum multipath error occurs if the multipath signal is of equal power to the line-of-sight signal and the multipath delay $\tau_m = 0$. This is a rather unlikely case for normal propagation conditions like free view to the sky, under canopy or in a rural environment. Even more unlikely is the case of the multipath signal being stronger than the line-of-sight signal, in which case the receiver may lock onto the multipath signal and all the information it gathers is unrelated to the line-of-sight signal.

If α is the ratio of the multipath signal amplitude to the line-of-sight amplitude, then

$$|\epsilon_{mp'}| < \frac{\sin^{-1}(\alpha|R(\tau_m)|)}{2\pi} \lambda \quad (14.80)$$

and carrier-phase multipath is often irrelevant for many applications, apart from real-time kinematic (RTK) positioning.

The carrier-phase noise is modeled as a zero mean Gaussian random variable, whose variance $\sigma_{n'}^2$ is to a large extent independent on the used modulation scheme, and given by

$$\sigma_{n'}^2 = \frac{\lambda^2}{4\pi^2} \frac{B_{PLL}}{C/N_0} \left(1 + \frac{1}{TC/N_0}\right). \quad (14.81)$$

The noise is usually at the order of 1 mm or below and virtually irrelevant for many applications.

14.6.3 Doppler

Between tracking loop updates, the carrier phase (14.74) increases linearly. The carrier-phase increase between two samples equals $2\pi(f_{IF} + \hat{f}_d)T_s$. Usually this increase is stored in one NCO variable (including f_{IF}). Within a simple receiver implementation, the Doppler frequency \hat{f}_d can be extracted straightforwardly from this rate. Doppler measurements can be done regardless of whether phase lock is achieved or not (provided that code and frequency lock has been achieved, i.e., the channel is actually tracking the signal).

The model for the Doppler frequency derives from the carrier-phase model (14.79) via a time derivative. By denoting time derivatives with a dot over the symbol, for example $dx/dt = \dot{x}$, then the Doppler observation equation is given by

$$\hat{f}_d = -\frac{1}{\lambda} [\dot{\rho} - c(\dot{dt}_r - \dot{dt}^s) + p_{T''} + \epsilon_{mp''} + \epsilon_{n''}], \quad (14.82)$$

with $\lambda = c/f_L$ being the carrier wavelength.

For most GNSS applications, atmospheric delays and receiver/hardware delays do not contribute to the Doppler, as they vary only slowly with time. However, if highly accurate velocity estimates are required, then atmospheric delay variations on the order of maximal few millimeters per second need to be accounted for [14.33].

The most significant contributors to the Doppler are the line-of-sight velocity $\dot{\rho}$ and the receiver clock drift \dot{dt}_r . The satellite clock drift \dot{dt}^s is very small.

The transient error is denoted as $p_{T''}$ and is caused by the inability of the FLL or PLL to instantaneously follow the variations in the line-of-sight velocity.

Additionally, Doppler measurements are affected by multipath errors $\epsilon_{mp''}$, but they are considered to be rather small (similar as for carrier-phase multipath).

The Doppler noise is bias free and its variance $\sigma_{n''}^2$ is independent of the modulation scheme and carrier frequency,

$$\sigma_{n''}^2 = \frac{1}{4\pi^2 T^2} \frac{4B_{FLL}}{C/N_0} \left(1 + \frac{1}{TC/N_0}\right). \quad (14.83)$$

Using longer coherent integration times T , precise Doppler measurements are possible.

Generally speaking, timely integrated Doppler measurements represent the relative line-of-sight variation

with a higher accuracy than code pseudoranges (of course code pseudoranges give an absolute measurement and not only a variation). The terminology of integrated Doppler is also used by some receiver manufacturers for their method to derive the Doppler as time differenced carrier-phase measurements.

Even more sophisticated receivers use correlator values themselves and interpolation to provide more precise Doppler measurements. More specifically, the Doppler can be derived by fitting a polynomial to the correlator values

$$\hat{\alpha}_1 = \arg \min_{\alpha_0, \alpha_1, \alpha_2, \dots} \sum_k \left(\frac{2\pi\varphi(k)}{\lambda} + \arctan \frac{Q_{P,k}}{I_{P,k}} - \alpha_0 - 2\pi\alpha_1 T k - \alpha_2 k^2 T^2 + \dots \right)^2. \quad (14.84)$$

The linear term $\alpha_1 = \hat{f}_d$ of the fitted polynomial is the Doppler frequency at the $k = 0$ point of the considered interval. For pilot signals, a four-quadrant arctan can be used. The Doppler values obtained in this way are free from any transient PLL or FLL errors and solely depend on the prompt correlation values in the considered interval. The higher the considered polynomial order, the noisier the Doppler measurements; the polynomial order can be seen as a replacement for the loop bandwidth.

14.6.4 Signal Power

The fourth observable that is generated by the receiver for each tracked signal is the estimated received signal power. In contrast to the other three observables, the signal power is not based on NCO variables, but can

readily be derived from (14.21)

$$\widehat{C/N_0} = \frac{I^2 + Q^2 - 2}{2T}. \quad (14.85)$$

The signal power estimates can be averaged over all prompt I/Q correlator values that are available during a measurement interval. For an integration time T of 20 ms, these are 50 values assuming a measurement rate of 1 Hz. Averaging power estimates increases the precision of the estimates, which is important for low signal power values. High sensitivity receivers often average over 5–10 s to detect the presence of low power 10–15 dB-Hz signals.

The signal power estimate $\widehat{C/N_0}$ is affected by thermal noise but is otherwise an unbiased estimate. The variance of the estimate is independent of the true C/N_0 if expressed in natural units and not if expressed in decibels.

Multipath signals influence $\widehat{C/N_0}$: constructive interference increases the value, destructive interference reduces the value. For static receivers, multipath can usually be well identified by a look at a C/N_0 time series as the fluctuations have a similar spectral character as the carrier-phase multipath. If the C/N_0 time series shows a variation with a period of for example 300 s, then the carrier-phase multipath error will have the same periodicity of 300 s. Even the amplitude of the C/N_0 fluctuations and the carrier-phase multipath are correlated but it is quite difficult to establish a precise quantitative relationship.

Transient errors in the code tracking loop decrease the signal power estimate as $R(\delta\tau)$ will then be smaller than 1. Similarly, Doppler frequency transient errors cause $\text{sinc}(\delta f_d T)$ to be smaller than 1 and also reduce the C/N_0 estimate.

14.7 Advanced Topics

The code correlation technique based on the full knowledge of the PRN code provides all components of signal parameters of interest to be used in navigation but cannot be available to the P(Y)-code tracking in dual-frequency receivers. Without knowledge of the Y-code, dual-frequency high-precision civilian purpose receivers employ codeless or quasicodeless techniques for the reconstruction of the unmodulated carrier wave in order to obtain precise carrier phase (and partly code) measurements on L2.

In addition, the recent success of digital signal processing technologies provides a powerful tool to implement a software-based GNSS receiver that has extremely easy configurable properties, thus leading

to many new processing schemes. It is ideally suited to many interesting applications, in order to improve signal tracking methods in modern digital GNSS receivers [14.34].

In this section we discuss advanced topics of signal processing technique focusing on a special case of signal processing in dual-frequency high-precision civilian receivers as well as recent progress in digital signal processing techniques of GNSS receivers.

14.7.1 Tracking of GPS P(Y)

Since the signal processing of the legacy GPS L1 and L2 signals requires both L1 CA- and L1/L2 P(Y)-codes

until now and the P(Y)-code is very long and officially unknown, only the L1 CA-code can be processed by the normal correlation technique. If the use of the P(Y)-code is guaranteed and officially available, both carriers on L1 and L2 can be easily reconstructed by the code correlation technique, enhancing the navigational performance. Without knowledge of the P(Y)-code, we can process it by applying codeless or quasicodeless techniques for the reconstruction of the unmodulated carrier wave. Currently, these techniques are widely used in dual-frequency GPS receivers for civilian survey purposes [14.35–38].

The first method is a squaring technique of the L2 P(Y)-code, which is based on autocorrelation of the received signal with itself to remove all modulations on L2. A 180° phase shift during modulation is equivalent to a change in the sign of the resultant signal, so the modulations will be removed by a squaring operation. However, the main drawback of this technique is that the satellite clock and orbit information, if the navigation data bits are modulated on the signal component, are lost in the process, and the signal-to-noise ratio is substantially degraded by 30 dB due to the squaring process. Also, the squaring operation results in the unmodulated carrier with the twice the original frequency, that is half the wavelength, in which the ambiguity resolution becomes more difficult [14.35, 36].

The second method is to use the cross correlation of L1 and L2. This technique is based on the fact that the unknown Y-code is identical on both carriers so that the cross correlation of the L1 and L2 signals is possible. The small delay due to the frequency-dependent propagation of an electromagnetic wave through the ionosphere is measurable as a variable. Therefore, the L2 P(Y)-code pseudorange and its carrier-phase measurements can be derived based on the cross-correlation output and the L1 CA-code pseudorange measurement. Full cycles of the L2 carrier can be retrieved. Due to the fact that the L1 P(Y)-code signal has twice the power of the corresponding L2 P(Y)-code signal, the cross correlation of the L1 and the L2 signal has an improvement of 3 dB in terms of the tracking threshold compared to the squaring of the L2 signal. However, compared to the code correlation technique, this method has a 27 dB degradation in the signal-to-noise ratio [14.35, 37].

The third method is so-called Z-tracking, which is an improved quasicodeless technique. This technique requires the P-code at a receiver to correlate separately with L1 and L2 signals with enough integration interval by a low-pass filter. The encryption signal bit (W-bit) is estimated separately in each frequency, which is then fed to the other frequency to remove the encryption

code from the signal. In this way, the code ranges and full wavelength L1 and L2 carrier phases are obtained. However, in comparison to the code correlation technique, this method results in a 14 dB of the signal-to-noise ratio degradation [14.35, 37].

Besides these, several other techniques were reported in the society but it is supposed that such P(Y)-code tracking methods will not be commonly used in the long-term future due to the arrival of new open GNSS signals.

14.7.2 Generic Data/Pilot Multiplexing Approach

The presence of data bits in a navigational signal significantly degrades the performance of signal tracking because it limits the coherent integration time, which should be long enough for a high sensitivity. The introduction of a pilot signal in addition to the data signal with a signal power split by multiplexing techniques is one of the major developments in modern GNSS signals. For example, GPS L1C and L5, and Galileo E1-OS and E5 use this technique [14.28–30, 39, 40].

After a lot of research works on the optimal power split between the data and pilot signals considering different GNSS applications, the equivalent split of total power between the data and pilot signals optimized for the 50 bit/s navigation data was finally proposed for GPS L5 [14.29, 41, 42]. An optimal power split at a much higher data rate was later discussed for SBAS L5 signals [14.43, 44], and the trade-off analysis between data rate and signal power split in GNSS signal design was well performed in [14.40]. Improvements from having a pilot channel and sharing the signal power comprise the following:

- The use of coherent tracking for a pilot channel may improve a somewhat poor tracking accuracy by eliminating the squaring loss of Costas tracking for a data channel.
- The use of pilot-only (50%) tracking may provide a C/N_0 gain of 6 dB by doubling the PLL threshold while only moderately increasing the PLL thermal noise jitter as compared with that of data-only (50%) tracking.
- The use of combined tracking may provide a C/N_0 gain of 3 dB over pilot-only tracking and with fewer tracking errors.
- The use of pilot channels may allow direct measurement of the carrier phase without the need to decode the navigation message preamble (see Sect. 14.7.3 for more details).

14.7.3 Combined Processing of Data and Pilot Signals

Two signal channels (data/pilot) multiplexed at a single satellite signal allow several combinations of signal tracking for the performance improvement. The followings are three examples of the available combination of data/pilot tracking:

- Independent data and pilot tracking
- Pilot-only tracking and aiding of data demodulation
- Combined data/pilot tracking.

Basically, the pilot channel provides benefits such as a high sensitivity in terms of a PLL thermal noise jitter and a wider available region of C/N_0 . The use of a pure PLL for the pilot channel, instead of the Costas loop used for the data channel, improves sensitivity as well as reliability, thereby gaining 4–9 dB in the carrier tracking [14.45, 46]. Here, an important question arises: Is a pilot channel alone sufficient for signal tracking? Or, is a data channel required in combination? The combined data/pilot tracking in an optimal manner will be beneficial in terms of high sensitivity compared to data-only tracking in Costas loop or pilot-only tracking in pure PLL [14.40]. However, due to the limited linearity of the pull-in-range of the Costas loop discriminator compared to that of the pure PLL discriminator, a proper combining method should be chosen depending on C/N_0 .

Figure 14.19 shows the block diagram of the signal tracking loop with a data/pilot combining algorithm. In order to effectively use all the signal power originally transmitted by a satellite for tracking, while the data-only or pilot-only tracking merely use half, the data and pilot signal components should be processed in combination. For this, a linear combination of a pure PLL discriminator and a Costas loop can provide an optimal

combining algorithm when based on the appropriate weighting coefficient for each channel. An unbiased estimate of the carrier-phase error (combined discriminator output) is given by [14.45]

$$\delta\phi_c = \alpha_D \delta\phi_D + \alpha_P \delta\phi_P. \quad (14.86)$$

Here, $\delta\phi_c$, $\delta\phi_D$, and $\delta\phi_P$ are the carrier-phase discriminator outputs from the combined, data, and pilot channels respectively. The weighting coefficients

$$\alpha_D = \frac{\sigma_P^2}{\sigma_D^2 + \sigma_P^2} \quad \alpha_P = \frac{\sigma_D^2}{\sigma_D^2 + \sigma_P^2}, \quad (14.87)$$

of the data and the pilot channels obey the relation $\alpha_D + \alpha_P = 1$ and can be obtained from the 1- σ PLL thermal noise jitter of the data and pilot channels respectively.

In the case of a 50-50 power split between data/pilot channels, the combined tracking may have a C/N_0 gain of 3 dB over the pilot-only tracking when C/N_0 is high enough. However, the pilot-only tracking is beneficial in low C/N_0 region due to its threshold efficiency. Therefore, a logic to select combined tracking or pilot-only tracking based on the C/N_0 estimate should be applied [14.40].

14.7.4 Combined Processing of Code and Carrier

In order to estimate signal parameters in (14.4), a classical tracking method widely uses single-input-single-output (SISO) DLL/PLL/FLL separately for code and carrier tracking as described in Sect. 14.4.1. The well-known DLL/PLL is equivalent to the special case of the linear quadratic Gaussian (LQG) optimal controller that is the combination of the Kalman filter (KF) and the output state feedback controller. Based on this, the

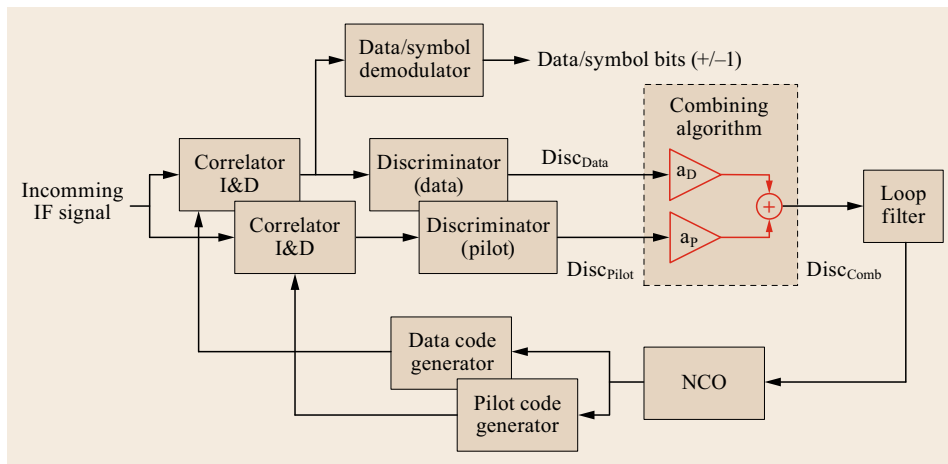


Fig. 14.19 Block diagram of signal tracking loop with a data/pilot combining algorithm

code and carrier tracking problem can be considered as a multiple-input-multiple-output (MIMO) control problem, which combines the code and carrier tracking altogether by means of the state feedback [14.47].

For a single satellite signal, a filter equation for a combined DLL/PLL/FLL including code and carrier NCOs based on concurrent code/carrier tracking is given in a state space form [14.1, 48] by the expression

$$\mathbf{x}_{k+1} = \mathbf{F}\mathbf{x}_k + \mathbf{L}e_k. \quad (14.88)$$

The statevector

$$\mathbf{x} = [A, \tau, \phi, f_d, \dot{f}_d]^\top,$$

comprises the amplitude A , the code delay τ , the carrier phase ϕ , the Doppler shift f_d , and the Doppler rate \dot{f}_d . The transition matrix

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & sT & 0 \\ 0 & 0 & 1 & T & 0 \\ 0 & 0 & 0 & 1 & T \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

is used to propagate the state across the update interval T , and the gain

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \omega_\tau T & 0 & 0 \\ 0 & 0 & 2.4w\omega_\phi T & 0 \\ 0 & 0 & 1.1w\omega_\phi^2 T & \sqrt{2}(1-w)\omega_{\dot{f}_d} T \\ 0 & 0 & w\omega_\phi^3 T & (1-w)\omega_{\dot{f}_d}^2 T \end{bmatrix}$$

determines the state correction for a given vector

$$\mathbf{e} = [\delta A, e_\tau, e_\phi, e_{\dot{f}_d}]^\top$$

of DLL, PLL, and FLL discriminator outputs, additionally including the signal amplitude difference.

In the above equations, $\omega_\tau, \omega_\phi, \omega_{\dot{f}_d}$ are the natural frequencies of the DLL/PLL/FLL and $w \leq 1$ is the weighting factor of the PLL relative to the FLL. Finally, $s = R_c/f_L$ denotes a scale factor used in the carrier-aided DLL for converting the carrier Doppler in Hertz to the code Doppler in chips per second. It is given by the ratio of the code chipping rate R_c (in chips per second) and the carrier frequency f_L (in Hz), since the Doppler effect on the signal is inversely proportional to the wavelength of the signal.

It is noted that if the phase error input of the PLL becomes zero or w is set to zero, the filter operates as a pure FLL and vice versa. If both phase and frequency error inputs do not sufficiently converge to zero, the use

of w makes it possible that the FLL assists the PLL to handle the phase error of $\pm 90^\circ$ and the PLL helps the FLL to provide a correct frequency error at 180° phase reversals.

It is also noted that in (14.88) a first-order carrier-aided DLL with narrow bandwidth for code tracking, and a combination of a second-order FLL and a third-order PLL for the carrier phase and Doppler tracking are specially adopted.

14.7.5 Carrier Tracking Kalman Filter

The traditional signal tracking architecture can be substituted by a Kalman filter (KF). A KF is equivalent to a digital phase-locked-loop (DPLL) with a time-invariant proportional integrator (PI) controller, which includes an integrator in the closed-loop feedback, whereas the KF has time-varying coefficients recursively computed by the KF equations in a form of a Kalman gain matrix [14.49]. The constant gain matrix in the DPLL is equivalent to the Kalman gain in the steady state, which can be obtained by numerical computation of the Kalman equations until convergence [14.50], or by the closed-form expressions [14.51].

Several different forms of KFs for signal tracking have been proposed especially for deep coupling architectures of a GNSS receiver. They can be grouped into several different approaches depending on how the measurement equations are constructed, or how the combination of state variables is made. For example, baseband I and Q components can be directly used as measurements [14.47, 52]. Also, nonlinear discriminator outputs are able to be used as measurements for a loop filter excluding NCO (i. e., error state KF) [14.53, 54] or measurement residual for an overall tracking loop including NCO (i. e., direct state KF) [14.1]. Both are equivalent to each other in terms of the loop filter transfer function [14.55].

In the following we consider the case of a carrier tracking KF, but the discussion is similar for code tracking and can also be extended to combined code/carrier tracking.

Based on the practical assumption that the carrier measurement residuals can be directly obtained from nonlinear discriminator outputs (i. e., $\tilde{\mathbf{z}} = [\delta\phi, \delta f_d]^\top$), the signal dynamics and measurement models for a signal tracking KF to substitute the carrier tracking loop described in the previous section are given in similar way as for conventional KF design [14.55]

$$\mathbf{x}_{k+1} = \mathbf{F}\mathbf{x}_k + \mathbf{w}_k, \quad (14.89)$$

$$\mathbf{z}_k = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k, \quad (14.90)$$

where $\mathbf{x}_k = [\phi, f_d, \dot{f}_d]_k^\top$ is the state vector at time k , \mathbf{z}_k is the measurement vector, \mathbf{w}_k is the process noise vector, and \mathbf{v}_k is the measurement noise vector. The noise vector \mathbf{w}_k and \mathbf{v}_k should satisfy the well-known additive white Gaussian noise assumption [14.56]. The state transition matrix \mathbf{F} and the design matrix \mathbf{H} are given by

$$\mathbf{F} = \begin{bmatrix} 1 & T & \frac{T^2}{2} \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

The measurement update equation of the signal tracking KF can be written in a state-space form as

$$\hat{\mathbf{x}}_k = \mathbf{F}\hat{\mathbf{x}}_{k-1} + \mathbf{K}\tilde{\mathbf{z}}_k, \quad (14.91)$$

where $\mathbf{K} = [k_{ij}]$ is a 3×2 Kalman gain matrix, which can be computed by the well-known Kalman filter equations

$$\mathbf{K} = \mathbf{P}_k^- \mathbf{H}^\top (\mathbf{H} \mathbf{P}_k^- \mathbf{H}^\top + \mathbf{R}_k)^{-1} \quad (14.92)$$

with the propagation and update equations for the covariance matrices

$$\mathbf{P}_{k+1}^- = \mathbf{F} \mathbf{P}_k^- \mathbf{F}^\top + \mathbf{Q}_k, \quad (14.93)$$

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K} \mathbf{H}) \mathbf{P}_k^-, \quad (14.94)$$

where $\hat{\cdot}$ and $^-$ denote the notations of a posterior estimate and of a priori estimate respectively, \mathbf{P} is a 3×3 state covariance matrix, \mathbf{R} is a 2×2 measurement noise covariance matrix, and \mathbf{Q} is a 3×3 process noise covariance matrix.

In fact, KFs require exact knowledge of the noise statistics in its initialization and tuning processes, which are to set the initial state vector (\mathbf{x}_0) and its error covariances (\mathbf{P}_0 , \mathbf{Q} , \mathbf{R}). This should be done empirically by fully employing a priori information of the system and the circumstances in which the KF operates. For example, in the signal tracking KF in a GNSS receiver, the coarse Doppler estimates $\hat{f}_{d,0}$ and its error range from the acquisition can be used for the initialization of $\hat{\mathbf{x}}_0$ and \mathbf{P}_0 . Also, an empirical knowledge on the noise statistics of discriminator outputs can be used for tuning \mathbf{R} . Therefore, tuning of \mathbf{Q} effectively determines the overall behavior of the signal tracking KF. Note that the uncertainty of the system dynamics model is dependent on the user dynamics, whereas that of the measurement model is governed by C/N_0 . Therefore, an adaptive scheme can be employed to adjust the equivalent noise bandwidth by regularly testing the process and mea-

surement noise statistics in response to changes in the signal dynamics as well as C/N_0 [14.57, 58].

Note that the signal tracking KF example described here is for the carrier-phase tracking loop, which uses both the PLL and FLL discriminator output for its measurement residuals, but can be also used for the PLL-only, the FLL-only, the DLL or their combined tracking with the relevant modifications of the architecture.

14.7.6 Vector Tracking

The concept of vector-based tracking technology, which is known as one of the most advanced signal processing architectures of modern digital GNSS receivers, was originally proposed in the early 1980s [14.59–63]. The most commonly cited reference for this appeared in the mid-1990s [14.64, 65]. Figure 14.20 shows an internal block diagram of a vector-tracking architecture.

The vector-tracking combines the signal tracking and navigation processing into one algorithm by a global optimum feedback through a line-of-sight projector, whereas a conventional scalar-tracking tracks each satellite's signal independently followed by a separate navigation solution processing [14.64, 65]. This means that in the vector-tracking mode individual tracking loops are eliminated and effectively replaced by a global feedback path from a navigation filter to each local channel to make a loop closure. Therefore, it enables efficient channel interactions, which represents the use of the redundant number of available satellites and their geometry in tracking a single satellite. It provides many benefits over the conventional scalar tracking, such as increased antijamming robustness, a higher signal tracking sensitivity, a better robustness to receiver dynamics, the ability to bridge signal outages and immediately reacquiring blocked signals [14.66]. Its primary drawbacks, however, are increased processing load and complexity, and instability of tracking channel by presence of a fault in one channel [14.1, 67, 68].

In fact, the vector tracking is originally based on the fact that the navigational solution parameters (i. e., position and time) of a GNSS receiver are not directly observable from the received signals. They are obtained through a two-layered nonlinear process: the first layer extracts the signal parameters of interest such as code delay, carrier phase, and Doppler from the received signals that are easily converted to range and range-rate measurements, and then the second layer calculates the navigational solution parameters by using these measurements. In other words, the received signal is a nonlinear function of the user's navigational solution parameters. Therefore, the vector-tracking architecture is implemented by closing the loop all the

way back to the signal correlators instead of having two separate shorter loops as in the conventional scalar-tracking architecture.

Theoretically the vector-tracking architecture has only one position estimator to calculate the navigational solution parameters directly from the received signals [14.53, 54, 69]. However, this approach has a serious drawback in a real implementation due to the physical limits of processing power; the signal processing channel should operate at a relatively higher rate than the navigation processing part and CPU limitations can mean that the navigation processing part cannot catch up with such a high operation speed. Moreover, the asynchronous property of each local signal tracking channel can be problematic [14.1, 70].

For this reason, a two-step filtering approach that has high-rate internal tracking loops and a low-rate navigation process (Fig. 14.20) is widely used [14.70–74]. Fundamentally, this approach is equivalent to the determination of efficient optimal parallel processing structures for channels that provide optimal estimates of the global system states for the navigational solution parameters. An appropriate error covariance at each layer should also be provided to the next layer for system optimality. A first look at this type of systems seems to turn back to the conventional scalar-tracking system, but it is noted that the main advantages of vector-tracking stated before are obtained from the efficient channel interaction performed by the global feedback, not from the construction of nonlinear measurement equations of received signals in terms of navigational solution parameters [14.1].

Therefore, in the vector tracking, two techniques, that is the efficient use of signal tracking KF and the channel interaction by the global loop closure, play an important role in achieving the performance im-

provement compared to the conventional scalar tracking [14.1, 70].

Another efficient method that links the high-rate internal tracking to the navigational solution is to first compute high-rate code pseudoranges and Doppler measurements and then model them during the measurement interval (e.g., every 1 s) by a polynomial as described in Sect. 14.6.3. The polynomial is then evaluated at the navigation update epoch for PVT computation. Practical results of this *polyfit* vector tracking method can be found in [14.75].

The outputs of signal tracking KFs in a local channel bank are then inputted into the navigation processor to solve the navigational equation as well as to produce feedback values for the NCO command in the individual local channels through a line-of-sight projector. The line-of-sight projector for the i th local channel, which is to project the navigational solution vector onto the line-of-sight domain of each channel, is implemented via a design matrix of the navigation solution and the predicted state variables. Without loss of generality it is expressed by the error state vector expression [14.70]

$$\begin{bmatrix} \delta \hat{\rho}_i \\ \delta \hat{\rho}_{\dot{i}} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{H}}_{\rho_i} \\ \hat{\mathbf{H}}_{\dot{\rho}_i} \end{bmatrix} \delta \hat{\mathbf{x}}_N, \quad (14.95)$$

where $\delta \hat{\rho}_i$ and $\delta \hat{\rho}_{\dot{i}}$ are the residual of predicted pseudorange and range rate respectively, $\hat{\mathbf{H}}_{\rho_i}$ and $\hat{\mathbf{H}}_{\dot{\rho}_i}$ are the partials of the modeled pseudorange and range rate with respect to the navigation solution, and $\delta \hat{\mathbf{x}}_N$ is the residual of the navigation solution vector (i.e., global solution). The residual of predicted pseudorange and range rate, that is the projected residual of the navigation solution vector is converted to NCO commands, and then fed into the code/carrier generators to control replica signals in the next step.

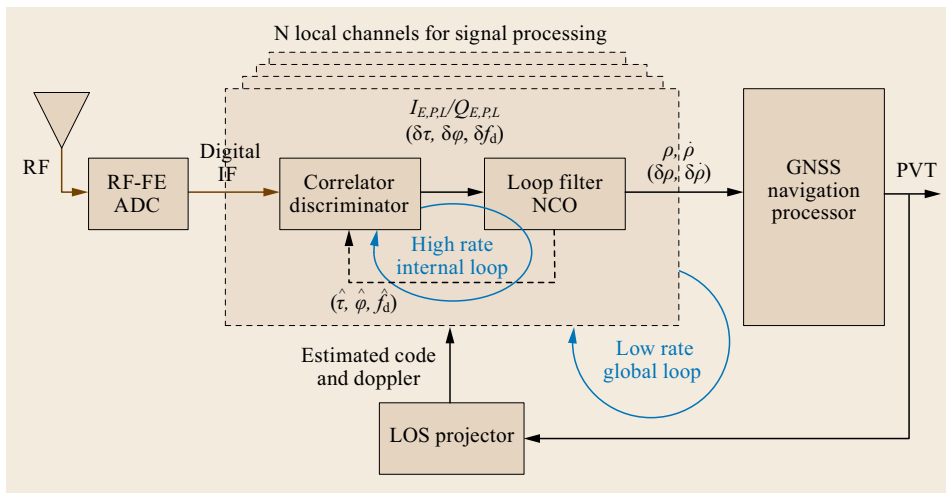


Fig. 14.20
Internal block
diagram of
a vector-tracking
architecture

Since the predicted pseudorange is less accurate than the corresponding carrier phase, it is critical when integrating the PLL into the vector-tracking architecture. If carrier-phase measurements are not required for positioning, the carrier phase should not be calculated from the projected pseudoranges. Instead, the carrier phase is freely running and determined by integrating the feedback Doppler [14.70].

It is necessary that all the channel measurements in a navigation processor should be synchronous, in order to achieve meaningful observables with a common clock bias, but that the start and stop boundaries of integration and dump process in a bank of channels are asynchronous. The meaningful observables having a common clock bias in a bank of channels can be obtained by propagating the pseudorange and range rate at a channel dump time to the receiver's fundamental time frame that is common to all the channels. This is

performed in a NCO by using a state transition matrix for the corresponding time offset in the individual channel [14.70].

The benefit of this channel interaction effect was well reported analytically [14.68] and numerically [14.76] under the assumption of constant pseudorange residuals. Also, the benefits of various signal tracking KFs were analyzed by means of numerical simulations [14.1]. Overall, it is known from previously performed field tests and analyses that the vector tracking can provide about 7 dB gain in terms of C/N_0 compared to the conventional scalar tracking, where around 3 dB gain can be obtained from the efficient use of signal tracking KF and the other 4 dB gain from the channel interaction [14.1, 77]. This gain makes it possible to bridge a certain level of faded signals that very often occurs in urban canyons [14.70, 78].

References

- 14.1 J.H. Won, D. Doetterboeck, B. Eissfeller: Performance comparison of different forms of Kalman filter approaches for a vector-based GNSS signal tracking loop, *Navigation* **57**(3), 185–199 (2010)
- 14.2 A.J. Van Dierendonck: GPS receivers. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996) pp. 329–407
- 14.3 O. Julien, C. Macabiau, M.E. Cannon, G. Lachapelle: ASPECT: Unambiguous sine-BOC(n, n) acquisition/tracking technique for navigation applications, *IEEE Trans. Aerosp. Electron. Syst.* **43**(1), 150–162 (2007)
- 14.4 P. Fine, W. Wilson: Tracking algorithm for GPS offset carrier and signals, *Proc. ION NTM 1999*, San Diego (ION, Virginia 1999) pp. 671–676
- 14.5 B. Barker, B.C. Barker, J.W. Betz, J.E. Clark, J.T. Correia, J.T. Gillis, S. Lazar, K.A. Rehborn, J.R. Straton et al.: Overview of the GPS M Code Signal (MITRE, 2002) pp. 1–8, Technical Paper
- 14.6 R.L. Fante: Unambiguous tracker for GPS binary offset carrier signals, *Proc. ION-AM-2003*, Albuquerque (ION, Virginia 2003) pp. 141–145
- 14.7 T. Pany: *Navigation Signal Processing for GNSS Software Receivers* (Artech House, Norwood 2010)
- 14.8 S.M. Kay: *Fundamentals of Statistical Signal Processing: Detection Theory* (Prentice Hall, Englewood Cliffs 1998)
- 14.9 T. Pany, E. Göhler, M. Irsigler, J. Winkel: On the state-of-the-art of real-time GNSS signal acquisition: A comparison of time and frequency domain methods, *Proc. Int. Conf. Indoor Position. Indoor Navig. (IPIN)*, Zurich (2010) pp. 1–8
- 14.10 E.D. Kaplan, C.J. Hegarty: *Understanding GPS – Principles and Applications*, 2nd edn. (Artech House, Boston/London 2006)
- 14.11 C.-H. Chiou, C.-W. Huang, K.-A. Wen, M.-L. Wu: A programmable pipelined digital differential matched filter for DSSS receiver, *IEEE J. Selected Areas Commun.* **19**(11), 2142–2150 (2001)
- 14.12 C. Stoeber, F. Kneissl, B. Eissfeller, T. Pany: Analysis and verification of synthetic multicorrelators, *Proc. ION GNSS 2011*, Portland (ION, Virginia 2011) pp. 2060–2069
- 14.13 D. Borio, L. Camoriano, L. Lo-Presti: Impact of acquisition searching strategy on the detection and false alarm probabilities in a CDMA receiver, *Proc. IEEE PLANS*, San Diego (2006) pp. 1100–1107
- 14.14 J. Tal: On the pull-in range of phase-locked loops, *IEEE Trans. Commun.* **23**(3), 390–393 (1975)
- 14.15 J.H. Won, P. Pany, B. Eissfeller: Iterative maximum likelihood estimators for GNSS signal tracking, *Trans. IEEE Aerosp. Electron. Syst.* **48**(4), 2875–2893 (2012)
- 14.16 A.J. Van Dierendonck, P. Fenton, T. Ford: Theory and performance of narrow correlator spacing in a GPS receiver, *Navigation* **39**(3), 265–284 (1992)
- 14.17 R. Jaffe, E. Rehtin: Design and performance of phase-lock circuits capable of near-optimum performance over a wide range of input signal and noise levels, *IEEE Trans. Inf. Theory* **1**(1), 66–76 (1955)
- 14.18 J.H. Won: Studies on the Software-Based GPS Receiver and Navigation Algorithms, Ph.D. Thesis (Ajou University, Suwon 2004)
- 14.19 J.W. Betz: Binary offset carrier modulations for radionavigation, *Navigation* **48**(4), 227–246 (2002)
- 14.20 J.J. Spilker Jr.: GPS signal structure and theoretical performance. In: *Global Positioning System: Theory and Applications*, Vol. 1, (AIAA, Washington DC 1996) pp. 57–119
- 14.21 M. Irsigler, B. Eissfeller: PLL tracking performance in the presence of oscillator phase noise, *GPS Solution* **5**(4), 45–57 (2002)

- 14.22 D. Allan: Statistics of atomic frequency standards, Proc. IEEE **54**(2), 221–230 (1996)
- 14.23 C. Hegarty, M.B. El-Arini, T. Kim, S. Ericson: Scintillation modeling for GPS-wide area augmentation system receivers, Radio Sci. **36**(2), 1221–1231 (2011)
- 14.24 T.E. Humphreys, M.L. Psiaki, P.M. Kintner Jr, B.M. Ledvina: GPS carrier tracking loop performance in the presence of ionospheric scintillations, Proc. ION GNSS 2005, Long Beach (ION, Virginia 2005) pp. 156–167
- 14.25 J.H. Won, B. Eissfeller, T. Pany, J. Winkel: Advanced signal processing scheme for GNSS receivers under ionospheric scintillation, Proc. IEEE/ION PLANS 2012, Myrtle Beach (ION, Virginia 2012) pp. 44–49
- 14.26 N.I. Ziedan: *GNSS Receivers for Weak Signals* (Artech House, Norwood 2006)
- 14.27 T.-Y. Chiou, D. Gebre-Egziabher, T. Walter, P. Enge: Model analysis on the performance for an inertial aided FLL-assisted-PLL carrier-tracking loop in the presence of ionospheric scintillation, Proc. ION NTM 2007, San Diego (ION, Virginia 2007) pp. 1276–1295
- 14.28 Global Positioning Systems Directorate: Navstar GPS Space Segment/Navigation User Segment Interfaces, Interface Specification (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo 2013) IS-GPS-200H
- 14.29 Global Positioning Systems Directorate: Navstar GPS Space Segment/User Segment L5 Interfaces, Interface Specification (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo 2013) IS-GPS-705D
- 14.30 European GNSS (Galileo) Open Service Signal. In: *Space Interface Control Document*, OS SIS ICD, Iss. 1.2, Nov. 2015 (European Union, 2015)
- 14.31 M. Irsigler, B. Eissfeller: Comparison of multipath mitigation techniques with consideration of future signal structures, Proc. ION GPS 2003, Portland (ION, Virginia 2003) pp. 2585–2592
- 14.32 J.W. Betz, K.R. Kolodziejewski: Extended theory of early-late code tracking for a bandlimited GPS receiver, Navigation **47**(3), 211–226 (2000)
- 14.33 A. Wieser: *GPS Based Velocity Estimation and Its Application to an Odometer* (Shaker, Aachen 2007)
- 14.34 T. Pany, J.H. Won, G. Hein: GNSS software defined radio: Real receiver or just tool for experts?, Inside GNSS Mag. **1**(5), 66–76 (2006)
- 14.35 B. Hoffmann-Wellenhof, H. Lichtenegger, E. Wasle: *GNSS – Global Navigation Satellite Systems* (Springer, Wien 2008)
- 14.36 J. Ashjaee: An analysis of Y-code tracking techniques and associated technologies, Geod. Info Mag. **7**(7), 26–30 (1993)
- 14.37 J. Ashjaee, R. Lorenz: Precision GPS surveying after Y-code, Proc. ION-GPS-92, Albuquerque (ION, Virginia 1992) pp. 657–659
- 14.38 K.T. Woo: Optimum semi-codeless carrier phase tracking of L2, Navigation **47**(2), 82–99 (2000)
- 14.39 Global Positioning Systems Directorate: Navstar GPS Space Segment/User Segment L1C Interfaces, Interface Specification (Global Positioning Systems Directorate, Los Angeles 2013) IS-GPS-800D
- 14.40 J.H. Won, B. Eissfeller, A. Schmitz-Peiffer, J.-J. Floch, F. Zanier, E. Colzi: Trade-off between data rate and signal power split in GNSS signal design, Trans. IEEE Aerosp. Electron. Syst. **48**(3), 2260–2281 (2012)
- 14.41 T. Morrissey: *Forward Error Correction for GPS L5 Data* (RTCA SC159 WG1, London 1999) pp. 20–21
- 14.42 C. Hegarty, A.J. Van Dierendonck: Civil GPS/WAAS signal design and interference environment at 1176.45 MHz: Results of RTCA SC159 WG1 activities, Proc. ION GPS 1999, Nashville (ION, Virginia 1999) pp. 1727–1736
- 14.43 M. Tran, C. Hegarty, A.J. Van Dierendonck, T. Morrissey: SBAS L1/L5 signal design options, Proc. ION GPS AM 2003, Albuquerque (ION, Virginia 2003) pp. 507–517
- 14.44 M. Tran: Performance evaluations of the new GPS L5 and L2 Civil (L2C) signals, Navigation **51**(3), 199–212 (2004)
- 14.45 C. Hegarty: Evaluation of the proposed signal structure for the new civil GPS signal at 1176.45 MHz (MITRE Corporation, 1999), WN 99W0000034
- 14.46 O. Julien: Carrier-phase tracking of future data/pilot signals, Proc. ION GNSS 2005, Long Beach (ION, Virginia 2005) pp. 113–124
- 14.47 G.I. Jee: GNSS receiver tracking loop optimization for combined phase, frequency and delay locked loops, Proc. ENC-GNSS, Munich (2005)
- 14.48 J.H. Won, P. Pany, B. Eissfeller: Non-iterative filter-based maximum likelihood estimators for GNSS signal tracking, Trans. IEEE Aerosp. Electron. Syst. **48**(2), 1100–1114 (2012)
- 14.49 P.F. Driessen: DPLL bit synchronizer with rapid acquisition using adaptive Kalman filtering techniques, IEEE Trans. Commun. **42**(9), 2673–2675 (1994)
- 14.50 G.S. Christiansen: Modeling of a PRML timing loop as a Kalman filter, Proc. IEEE GLOBECOM, San Francisco, Vol. 2 (1994) pp. 1157–1161
- 14.51 A. Patapoutian: On phase-locked loops and Kalman filters, IEEE Trans. Commun. **47**(5), 670–672 (1999)
- 14.52 M.L. Psiaki, H. Jung: Extended Kalman filter methods for tracking weak GPS signals, Proc. ION GPS 2002, Portland (ION, Virginia 2002) pp. 2539–2553
- 14.53 D. Gustafson, J. Dowdle, K. Flueckiger: A deeply integrated adaptive GPS-based navigator with extended range code tracking, Proc. IEEE PLANS, San Diego (2000) pp. 118–124
- 14.54 D. Gustafson, D.E. Gustafson, J.R. Dowdle, J.M. Elwell jr: Deeply-Integrated Adaptive INS/GPS Navigator with Extended-Range Code Tracking, US Patent (Application) Ser., Vol. 6630904 B2 (2003)
- 14.55 J.H. Won, P. Pany, B. Eissfeller: Characteristics of Kalman filter approach for signal tracking loop of GNSS receiver, Trans. IEEE Aerosp. Electron. Syst. **48**(4), 3671–3681 (2012)
- 14.56 R.G. Brown, P.Y.C. Hwang: *Introduction to Random Signals and Applied Kalman Filtering with MATLAB Exercises and Solutions*, 3rd edn. (John Wiley, New York 1997)

- 14.57 J.H. Won, B. Eissfeller: A tuning method based on signal-to-noise power ratio for adaptive PLL and its relationship with equivalent noise bandwidth, *IEEE Commun. Lett.* **17**(2), 393–396 (2013)
- 14.58 J.H. Won: A novel adaptive digital phase-lock-loop for modern digital GNSS receivers, *IEEE Commun. Lett.* **18**(1), 46–49 (2014)
- 14.59 E.M. Copps, G.J. Geier, W.C. Fidler, P.A. Grundy: Optimal processing of GPS signals, *Navigation* **27**(3), 171–182 (1980)
- 14.60 J.W. Sennott: A flexible GPS software development system and timing analyzer for present and future microprocessor, *Navigation* **31**(2), 84–95 (1984)
- 14.61 J.W. Sennott, D. Senffner: Navigation Receiver with Coupled Signal-Tracking Channels, US Patent Application Ser., Vol. 5343209 (1992) Bloomington, IL
- 14.62 J.W. Sennott, D. Senffner: The use of satellite geometry for prevention of cycle slips in a GPS processor, *Navigation* **39**(2), 217–236 (1992)
- 14.63 J.W. Sennott, D. Senffner: Comparison of continuity and integrity characteristics for integrated and decoupled demodulation/navigation receiver, *Proc. ION GPS 1995*, Palm Springs (ION, Virginia 1995) pp. 1531–1537
- 14.64 J.J. Spilker Jr: Vector Delay Lock Loop Processing of Radiolocation Transmitter Signals, US Patent (Application) Ser., Vol. 5398034 (1995) Stanford Telecommunications Inc.
- 14.65 J.J. Spilker Jr: Fundamentals of signal tracking theory. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker Jr. (AIAA, Washington DC 1996) pp. 245–327
- 14.66 T. Pany, B. Eissfeller: Use of a vector delay lock loop receiver for GNSS signal power analysis in bad signal conditions, *Proc. IEEE PLANS*, San Diego (2006) pp. 893–903
- 14.67 D. Benson: Interference benefits of a vector delay lock loop (VDLL) GPS receiver, *Proc. ION AM 2007*, Cambridge (ION, Virginia 2007) pp. 749–756
- 14.68 M. Lashley, D.M. Bevely: What are vector tracking loops, and what are their benefits and drawbacks?, *Inside GNSS Mag.* **4**(3), 16–21 (2009)
- 14.69 J.M. Horslund, J.R. Hooker: Increase Jamming Immunity by Optimizing Processing Gain for GPS/INS Systems, US Patent Application 5983160 (1999) Raytheon Company
- 14.70 J.H. Won, B. Eissfeller: Implementation, test and validation of a vector-tracking-loop with the ipex software receiver, *Proc. ION GNSS 2011*, Portland (ION, Virginia 2011) pp. 795–802
- 14.71 P.Y. Kim, F.L. Orlando: GPS Navigation with Integrated Phase Tracking Filter, US Patent Application 7151486 B2 (2006) Lockheed Martin Corporation
- 14.72 A.S. Abbott, W.E. Lillo: Global Positioning Systems and Inertial Measuring Unit Ultratight Coupling Method, US Patent Application 6516021 B1 (2003) The Aerospace Corporation
- 14.73 A. Jovancevic, A. Brown, S. Ganguly, J. Noronha, B. Sirpatil: Ultra tight coupling implementation using real time software receiver, *Proc. ION GNSS 2004*, Long Beach (ION, Virginia 2004) pp. 1575–1586
- 14.74 E.J. Ohlmeyer: Analysis of an ultra-tightly coupled GPS/INS system in jamming, *Proc. IEEE/ION PLANS 2006*, San Diego (ION, Virginia 2006) pp. 44–53
- 14.75 T. Pany, N. Falk, B. Riedl, C. Stoeber, T. Hartmann, G. Stangl: Receiver technology, software receivers, an answer for precise positioning research, *GPS World* **23**(9), 60–66 (2012)
- 14.76 J.H. Won, B. Eissfeller: Effectiveness analysis of vector-tracking-loop in signal fading environment, *Proc. NAVITEC*, Noordwijk (2010) pp. 1–6
- 14.77 M.G. Petovello, G. Lachapelle: Comparison of vector-based software receiver implementations with application to ultra-tight GPS/INS integration, *Proc. GNSS 2006*, Fort Worth (ION 2006) (2006) pp. 1790–1799
- 14.78 S.J. Ko, B. Eissfeller, J.H. Won: Assessment of vector-tracking-loop performance under radio frequency interference environments, *Proc. ION GNSS 2012*, Nashville (ION, Virginia 2012) pp. 2333–2341

Multipath

15. Multipath

Michael S. Braasch

Multipath is the phenomenon whereby the signal from a satellite arrives at the receiver via multiple paths due to reflection and diffraction. These nondirect-path signals distort the received signal and cause errors in code and phase measurements. Differential techniques do not eliminate multipath and thus multipath is an important error source in high precision applications. The physical surroundings around the receiver's antenna dictate the multipath environment and thus cause significant differences for land, marine, airborne, and spaceborne users.

This chapter describes the multipath environment and presents models describing the impact of multipath on code and phase measurements. The influence of the type and rate of the broadcast code as well as the receiver architecture will be highlighted. Mitigation techniques based on receiver design will also be described along with the impact of receiver dynamics. Finally, a technique to measure multipath is described and its usage in evaluating static environments is discussed.

The goal of this chapter is to provide the reader with the tools to assess the impact of multipath on both the code and phase and to understand the performance improvements and limitations associated with various multipath mitigation techniques.

15.1	The Impact of Multipath	444
15.2	Characterizing the Multipath Environment	444
15.3	Multipath Signal Models	448
15.4	Pseudorange and Carrier-Phase Error	450
15.5	Multipath Error Envelopes	450
15.6	Temporal Error Variation, Bias Characteristics and Fast Fading Considerations	453
15.7	Multipath Mitigation	455
15.7.1	Multipath Mitigation via Antenna Placement	455
15.7.2	Antenna Type	456
15.7.3	Receiver Type	457
15.7.4	Measurement Processing	458
15.8	Multipath Measurement	459
15.8.1	Isolation of Pseudorange Multipath ...	460
15.8.2	Short-Delay Multipath	461
15.8.3	Multipath Repeatability	462
15.8.4	Measurement of Carrier-Phase Multipath	463
15.9	A Note About Multipath Impact on Doppler Measurements	466
15.10	Conclusions	466
	References	466

The signals broadcast by a given global navigation satellite system (GNSS) satellite at a particular moment in time will be received over a large fraction of the Earth's surface (specifically, the portion of the Earth that is *visible* to the satellite). The broadcast signal thus illuminates all objects in the vicinity of a given GNSS receiver in addition to the GNSS receiver-antenna itself. The composite signal received is therefore a combination of the direct line-of-sight (LOS) signal and signals that are reflected and/or diffracted from nearby objects. The *undesired* non-LOS signals cause distur-

tion of the desired LOS signal. This, in turn, leads to tracking errors in the receiver and subsequently code and phase measurement errors.

The term *multipath* is derived from the phrase *multiple paths* but is conventionally used to describe the non-LOS signal (paths) that degrade the desired direct (path) signal. The first step in mitigating the multipath problem is to characterize the multipath environment. Following this, the impact of the multipath on the receiver processing and measurement generation can be determined. Mitigation strategies may then be effec-

tively implemented. Finally, techniques are needed to measure multipath error on GNSS measurements so that assessments can be made of the magnitude of the multipath problem for a given environment as well as to determine the efficacy of any mitigation strategies that are employed.

In order to avoid possible confusion, it should be noted that a phenomenon other than multipath can occur in dense urban canyon environments. In such places

it is not uncommon for the LOS signal to be completely blocked (a situation that is also referred to as *shadowing*). It is then possible, in some cases, for a GNSS receiver to track and form measurements on non-LOS signals. Resulting errors can be very large since the measurements are not tied in any way to the direct path signal. This non-LOS signal tracking problem is separate from, and should not be confused with, the multipath problem that is the subject of this chapter.

15.1 The Impact of Multipath

Multipath is just one of many error sources that affect GNSS accuracy and thus its impact must be characterized in the context of a given environment and application. Besides multipath, typical GNSS error sources include satellite clock and ephemeris error (that induce ranging errors on the order of 1–2 m), ionospheric delay (with ranging errors varying from 5 to 30 m if uncompensated), tropospheric delay (ranging errors varying from 2 to 20 m if uncompensated), and receiver noise (pseudorange errors on the decimeter to meter level and carrier-phase errors on the order of millimeters to centimeters). Pseudorange multipath errors can range as high as 100 m in the most severe conditions whereas carrier-phase multipath errors range from millimeters to centimeters.

The severity of the multipath impact is strongly dependent upon the GNSS application. For example, aircraft utilizing GNSS for enroute navigation can tolerate several hundred meters of position error and thus multipath is not a significant error source. Ships traversing open water are in a very similar situation as are satellites utilizing GNSS for orbit determination and station keeping. In general, any application that does not require the use of differential corrections (either

ground based or space based) will typically not be significantly affected by multipath.

However, any application that does require the use of code differential corrections will find multipath to be problematic. This is due to the fundamental principle of differential GNSS: Differential corrections only remove those error sources that are common both to the remote unit and the reference station (Chap. 12 and Chap. 31). Even for closely spaced receivers, multipath error will not, in general, be removed by differential corrections. The distinct difference in multipath error at physically separated receivers is due mainly to the short carrier wavelength of GNSS signals. Even small receiver separations are relatively large in terms of GNSS wavelengths.

For differential carrier-phase applications, multipath has two primary effects. Since initial ambiguity estimates are typically derived from pseudorange measurements, pseudorange multipath will force a large search space and thus cause ambiguity resolution to take longer than would be the case in a cleaner environment. Once ambiguities have been resolved, carrier-phase multipath will limit the accuracy of the determined baseline.

15.2 Characterizing the Multipath Environment

The general multipath environment is illustrated in Fig. 15.1. In addition to the desired direct path signal, there are three non-LOS signals impinging upon the antenna of the GNSS receiver. There are two reflected signals (one from the building on the right and one from the ground) and one diffracted signal. The reflected signals obey Snell's law wherein the angle of incidence equals the angle of reflection. Diffracted signals emanate from the edges of objects, such as the roof corner of the building on the left, and scatter in a wide range of angles.

An important principle can be drawn from Fig. 15.1. All multipath signals travel a longer path than the direct

signal. A given point on the direct path signal will therefore arrive at the antenna earlier in time than the same point on any of the multipath signals. In general there are five parameters that characterize a given multipath signal:

1. Relative delay
2. Relative amplitude
3. Relative phase
4. Relative phase rate
5. Relative polarization.

Relative refers to the direct path signal. Thus multipath with a relative amplitude of, say, –10 dB, means the am-

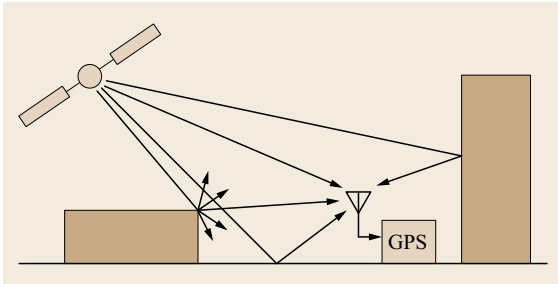


Fig. 15.1 General multipath environment

plitude of the multipath is 10 dB below the amplitude of the direct path signal. In some contexts, the *relative* is understood and thus is not stated explicitly.

For the case of simple reflections from flat surfaces, determination of relative delay can be aided through the use of *image theory* from electromagnetic theory. First consider the simple reflection case illustrated in Fig. 15.2. Note that since the distance from the antenna to the satellite is much, much larger than the distance between the antenna and the reflecting surface, the signal paths from the satellite are approximated as being parallel. Determining the relative delay of the multipath requires determining the extra path length traveled by the multipath. Although certainly a tractable problem, it is greatly simplified through the use of image theory as depicted in Fig. 15.3. An imaginary *image* antenna is placed at a distance h below the reflecting plane whereas the actual antenna is located at a height h above the plane. The path length from the satellite to the image antenna in Fig. 15.3 is the same path length as for the multipath depicted in Fig. 15.2. With image theory, however, determination of the additional multipath path length is a simple application of trigonometry.

The amplitude of the multipath is affected by the size, shape, and reflection coefficient of the multipath inducing surface. The primary factor affecting the amplitude of the reflected or diffracted signal is the reflection coefficient of the multipath-inducing surface. Detailed treatment of reflection coefficients will not be provided here but it must be noted that reflection coef-

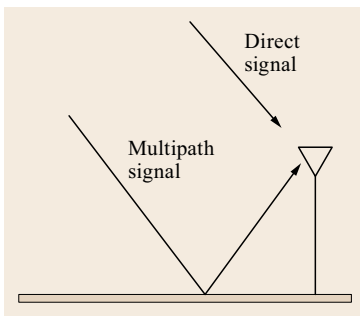


Fig. 15.2 Simple reflection case

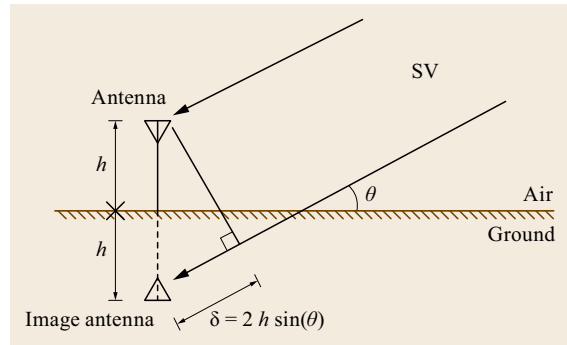


Fig. 15.3 The image antenna is located at a distance below the reflecting plane equal to the distance of the real antenna above the reflecting plane. The multipath relative delay is given by the extra distance that the signal has to travel to the image antenna versus the path to the real antenna. This may easily be obtained from the figure through simple trigonometry

ficients are a function of the angle of incidence of the signal and typically approach unity for near parallel incidence (e.g., very small elevation angles in the case of ground reflection). Thus, although dry soil will attenuate a reflected GNSS signal by approximately 10 dB for normal incidence (e.g., 90° elevation angle in the case of ground reflection), the amount of attenuation decreases as the elevation angle decreases. Table 15.1 lists approximate attenuation values for typical surface types for the L1 frequency with normal incidence. The table indicates that strong reflections are possible from wet soil, bodies of water, and tinted glass. As will be described later, surface roughness can effectively provide some attenuation to the reflected signal.

Following the theory of optics, significant signal reflection only occurs for objects large enough in size that they span a significant fraction of a cross-section of the

Table 15.1 Reflection coefficients and attenuation factors for common surfaces at normal incidence (elevation angle of 90°) at the L1 frequency (after [15.1–3])

Surface type	Reflection coefficient	Attenuation factor (dB)
Dry soil	0.268	−11.4
Moderate soil	0.566	−4.94
Wet soil	0.691	−3.21
Grassy field	0.334	−9.53
Asphalt	0.121	−18.3
Fresh water	0.800	−1.95
Sea water	0.811	−1.83
Glass	0.421	−7.51
Tinted glass	0.950	−0.446
Brick	0.345	−9.24
Concrete	0.404	−7.87

first Fresnel zone [15.4]. The first Fresnel zone is an ellipsoid about the LOS between a transmitting antenna and a receiving antenna consisting of all points each with a combined path length to the receiving and transmitting antennas that is one-half wavelength longer than the LOS distance. In the case of GNSS reflected signals, however, the receiving antenna is not the actual antenna but, rather, the image antenna. This is depicted in Fig. 15.4 for the same ground reflection scenario discussed earlier. The cross-section of the first Fresnel zone is observed to be an ellipse. It follows that if the satellites were directly overhead (i.e., normal incidence), the cross-sectional area would be a circle. The first-order approximation of the radius of this circle (known as the first Fresnel zone radius) is given by [15.5]

$$r = \sqrt{\frac{\lambda d_t d_r}{d_t + d_r}}, \quad (15.1)$$

where λ is the carrier frequency of the transmitted signal, d_t is the distance from the cross-section to the transmitter and d_r is the distance from the cross-section to the receiver. Note it is assumed that d_t and d_r are both much, much larger than λ .

For cases where d_t is much, much larger than d_r (such as GNSS), the denominator in (15.1) is approximated by $d_t + d_r \approx d_t$ and then (15.1) reduces to

$$r = \sqrt{\lambda d_r}. \quad (15.2)$$

For oblique incidence, the cross-section is an ellipse, not a circle. The length of the semimajor axis of this ellipse can be approximated by $a = r / \sin(\theta)$ where θ is the angle between the reflecting plane and the LOS to the satellite (Fig. 15.3). The Fresnel zone radius may be computed using (15.2) but note that in the oblique incidence case, the distance from the image antenna to the

point of reflection is given by the orthogonal distance of the antenna from the reflecting plane divided by the sine of θ .

As an example, if an antenna is 1 m above a reflecting surface and a satellite is directly overhead, the radius of the Fresnel zone cross-section (at the L1 frequency) is 0.436 m. For the same antenna height, the length of the semimajor axis of the cross-section for a satellite at 5° elevation angle is approximately 17 m. At a height of 10 m, the radius of the circular cross-section is 1.4 m and for a 5° elevation angle satellite, the semimajor axis length is approximately 54 m. For reflecting surfaces that are not horizontal, the same concepts apply but the actual calculations are somewhat more complex. In any case, it is clear that objects less than a wavelength in size are not significant sources of reflected energy. For oblique incidence, the reflecting surface needs to be dozens or even hundreds of wavelengths in size depending upon the orthogonal distance of the antenna from the reflecting plane.

Regarding reflecting obstacle shape, planar surfaces are the most problematic since they produce reflections that are focused in a particular region. Besides the surface of the earth itself (either the ground or bodies of water), planar surfaces abound particularly in urban canyon environments in the form of industrial and office buildings. A concept related to obstacle shape is surface roughness. A flat piece of metal is obviously an ideal reflector. However, what if one considers a field of tall grass (i.e., one that has not been mowed)? Although entire books have been written on the subject of electromagnetic scattering from rough surfaces [15.4], suffice it to say that rough surfaces do not produce pure specular reflection in a particular direction. Instead, each small locally flat surface produces a weak reflection in a direction consistent with Snell's law. This yields a broadly scattered field. In this context, the electromagnetic *roughness* of a surface is a function of the wavelength of the transmitted signal and the angle of incidence. A surface can be considered electromagnetically rough if it has variations in height that are on the order of a wavelength or more and in which peak to trough variations occur laterally with spacing also on the order of a wavelength. However, this characterization only holds for normal incidence (i.e., an elevation angle of 90° in the case of ground reflection). Even rough surfaces appear smooth for very small grazing angles (e.g., elevation angles in the case of ground reflection).

Reflections from electromagnetically smooth surfaces are referred to as *specular* reflections whereas reflections from rough surfaces are referred to as *diffuse* reflections. As an example, diffuse reflections of GNSS signals from a grassy field will be attenuated by approximately 2–4 dB (for elevation angles ranging from

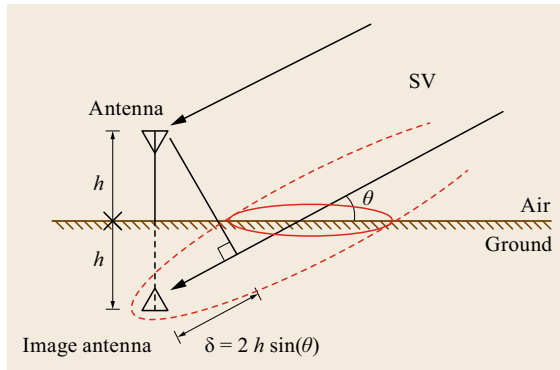


Fig. 15.4 Depiction of the ellipsoidal Fresnel zone along the path from the satellite to the image antenna

5–90°). Similarly, for dense brush/weeds, the attenuation ranges from 5 to 10 dB and for trees the attenuation ranges from 10 to 20 dB [15.6].

The relative phase of the multipath dictates whether it will constructively or destructively interfere with the direct signal. As will be described later, this results in a range of pseudorange and carrier-phase measurement errors that can occur for a multipath signal with a given delay and amplitude. The relative phase of the multipath is a function of the relative path delay and the reflection coefficient of the multipath-inducing surface. Since the reflection coefficient is, in general, a complex quantity, the reflected signal can experience an effectively instantaneous change of phase upon reflection.

The relative phase of the multipath can experience a nonzero time rate of change due to the relative motion of the transmitter, receiver, and multipath-inducing obstacle. As will be discussed later, GNSS receiver tracking loops will, to varying degrees, attenuate the multipath component of the received signal if its relative phase-rate is large compared to the tracking loop bandwidth. Thus, a GNSS-equipped automobile driving by a large building will not be affected as severely as when it is parked near the building.

The multipath phase-rate may be obtained by calculating the derivative of the multipath relative delay. As an example, consider the simple ground reflection scenario depicted in Fig. 15.3. The multipath relative delay is given by

$$\delta(t) = 2h(t) \sin[\theta(t)] . \quad (15.3)$$

It is instructive to consider two special cases: static and dynamic receivers. In the case of a static receiver (e.g., a ground reference station), the height is a constant and only the satellite elevation angle is changing. Conversely, in the case of a dynamic receiver (e.g., an aircraft descending to land), the satellite elevation angle can be considered constant (over short time intervals) whereas the receiver height is changing. Note the principle is the same if, say, a car is driving toward a large building; however the geometry is slightly more complex to analyze. For the static receiver case, the derivative of the relative delay is given by

$$\frac{d}{dt}\delta(t) = 2h \cos[\theta(t)] \frac{d\theta}{dt} . \quad (15.4)$$

Although precise rates-of-change of satellite elevation angle can be obtained using the broadcast ephemeris parameters for specific receiver locations, a rough rule-of-thumb can be obtained simply by recognizing that, at mid-latitudes, the main satellite pass of the day (i. e., satellite rise to satellite set) takes approximately 8 h. Thus roughly 180° of elevation are traversed

in 8 h which corresponds approximately to 10^{-4} rad/s. For a nominal satellite elevation angle of 30° and an elevation angle rate of 10^{-4} rad/s, (15.4) yields a multipath relative delay rate of $h \cdot 1.73 \cdot 10^{-4}$ m/s. This can be converted to a phase rate by dividing by the carrier wavelength. For the example considered, at L1 this yields $h \cdot 9.1 \cdot 10^{-4}$ Hz or, roughly, 1 mHz/m of antenna height. For typical ground reference station antennas, then, the error oscillations due to ground reflection are very low frequency. These frequencies are referred to in the literature as multipath phase rates or as *fading frequencies*.

Conversely, for dynamic receivers, significantly higher fading frequencies are possible. Over short time intervals (e.g., less than a minute), the satellite elevation angle can be considered a constant and the derivative of (15.3) yields

$$\frac{d}{dt}\delta(t) = 2 \sin(\theta) \frac{d}{dt}h(t) . \quad (15.5)$$

Again, this can be converted to a phase rate by dividing by the carrier wavelength. At L1, for example, a descent rate of 1 m/s results in a phase rate of approximately 0–10 Hz for satellite elevation angles ranging from 0 to 90°. For a specific application, consider an aircraft performing an approach to landing. For a typical 3° glide path angle and an approach speed of 100 knots, the vertical descent rate is approximately 2.7 m/s. For a nominal satellite elevation angle of 30°, (15.5) yields a relative delay rate of approximately 1.3 m/s and, at L1, a fading frequency of approximately 7 Hz. Note this is 3 orders of magnitude larger than the ground reflection fading frequencies generated for a static receiver. As will be discussed later, certain receiver architectures can significantly attenuate the so-called fast fading multipath.

The last parameter is the relative polarization of the multipath signal. GNSS signals are circularly polarized. A circularly polarized signal reflected from a perfect conductor (e.g., metal) will experience a polarization reversal. Thus if the incident wave is right-hand circularly polarized, the reflected signal will be left-handed. This is important since GNSS antennas are designed to receive the desired polarization and will attenuate the opposite polarization (typically on the order of 10 dB). For nonmetallic surfaces, however, the polarization of the reflected signal will be a mixture of both right- and left-handed yielding what is known as elliptical polarization. In this case the antenna will attenuate only that portion of the signal that is reverse polarized. Typically this is half the reflected field and thus approximately 3 dB of attenuation is achieved.

15.3 Multipath Signal Models

A received GNSS signal consisting of the LOS signal plus N reflected signals can be modeled as [15.7]

$$s(t) = \sum_{i=0}^N a_i(t) p[t - \tau_i(t)] \cos[\omega_0 t + \theta_i(t)] + \epsilon(t), \quad (15.6)$$

where $a_i(t)$ is the amplitude of the i -th component, $p(t)$ is the GNSS code modulation, $\tau_i(t)$ is the relative time delay of the i -th component, ω_0 is the nominal frequency of the LOS signal, $\theta_i(t)$ is the relative phase of the i -th component and $\epsilon(t)$ represents the noise. In this model, all other error sources are ignored. The LOS signal component corresponds to $i = 0$ and, without loss of generality, the LOS signal parameters are defined as: $a_0(t) = 1$, $\tau_0(t) = 0$, and $\theta_0(t) = 0$.

As described in Chap. 14, tracking of the code of most GNSS signals utilizes some form of delay-lock loop (DLL) also known as early-minus-late tracking. Tracking of the carrier involves processing a *prompt* channel that lies between the early and late channels. To understand the influence of multipath on both the code and carrier, we must first consider its impact on the correlation function. This is illustrated in Fig 15.5 for the simplified case of a binary phase shift keying (BPSK) signal with infinite bandwidth and a single multipath signal. The left illustration in the figure depicts the absence of multipath in which the main correlation peak is symmetrical about a lag of zero. In the center, the correlation peak associated with a single multipath signal is illustrated along with the direct path peak. These separate peaks are not observable in reality but are useful to consider separately for analysis purposes. On the right, the combination of the two is illustrated with the assumption that the multipath is *in phase* with the direct signal (e.g., the relative phase of the multipath is zero). This composite, distorted, correlation function is what the receiver must process for tracking. As will be described next, the asymmetry of the distorted correlation function results in tracking error that yields pseudorange

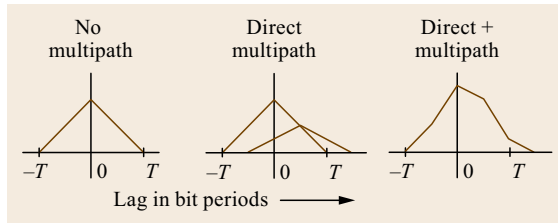


Fig. 15.5 Correlation function for combination of direct and reflected signal

range measurement error. The multipath also distorts the phase of the received signal and this leads to carrier-phase measurement error.

Figure 15.6 illustrates the formation of the DLL discriminator function (sometimes referred to as the S-function or S-curve) for the case of 1 chip early-to-late spacing. In the absence of multipath, the zero-crossing point of the discriminator function provides a tracking point that corresponds to the peak of the unshifted (referred to as *prompt*) correlation function. In the presence of multipath, however, the asymmetrical distortion of the correlation function induces a shift in the zero-crossing point. The amount of shift, when converted from chips to units of distance, is equal to the pseudorange error due to multipath.

There are two broad categories of DLLs. Coherent DLLs assume the carrier-tracking loop is locked in phase to the received signal. Noncoherent DLLs do not make this assumption and must, as a result, square the correlator outputs prior to forming the discriminator function. For the aforesaid general signal model, the coherent DLL discriminator function is given by [15.7]

$$S_{\text{coh}}(\tau) = \int_{-T}^{T} \sum_{i=0}^N a_i(t) \cos(\theta_i - \theta_c) \times \left[R\left(\tau - \tau_i + \frac{d}{2}\right) - R\left(\tau - \tau_i - \frac{d}{2}\right) \right] dt, \quad (15.7)$$

where θ_c is the phase of the composite received signal, $R(\tau)$ is the correlation function of the GNSS code for a lag value of τ , d is the early-to-late correlator spacing

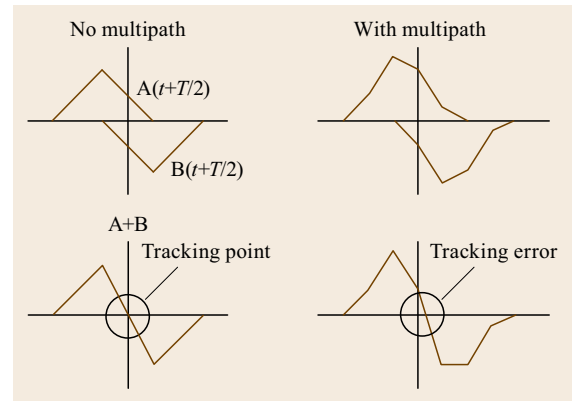


Fig. 15.6 Discriminator function for combination of direct and reflected signal

and T_{avg} is the time over which the correlator outputs are averaged before forming the discriminator function. Note the code tracking loop bandwidth is given by the one-sided noise bandwidth of the averaging [15.7]: $B_L = 1/(2T_{\text{avg}})$.

Although a variety of noncoherent discriminators exist, the simplest is the early-power-minus-late-power detector (Chap. 14). For this noncoherent DLL, the early and late correlator components are squared before they are differenced [15.7]

$$S_{\text{ncoh}}(\tau) = \int_t^{t+T_{\text{avg}}} \left| \sum_{i=0}^N a_i(t) R\left(\tau - \tau_i + \frac{d}{2}\right) e^{j\theta_i} \right|^2 - \left| \sum_{i=0}^N a_i(t) R\left(\tau - \tau_i - \frac{d}{2}\right) e^{j\theta_i} \right|^2 dt. \quad (15.8)$$

As described in Chap. 14, the DLL tracks the received signal by shifting its locally generated code so as to drive the discriminator function to zero. As shown in Fig. 15.6, in the presence of multipath the discriminator function zero-crossing point is erroneous. This zero-crossing point (or DLL tracking point of the composite signal) will be denoted by τ_c in this chapter.

The carrier-tracking loop tracks the phase of the composite received signal. This composite phase is given by the phase angle of the prompt correlator output. The prompt correlator output is given by the vector sum of the prompt correlation of each component given in (15.6) [15.7]

$$\theta_c = \arg \left[\int_t^{t+T_{\text{avg}}} \sum_{i=0}^N a_i e^{j\theta_i} R(\tau_c - \tau_i) dt \right], \quad (15.9)$$

where, as noted above, τ_c is the DLL code tracking error due to the multipath. Since the phase of the LOS signal is zero by definition, θ_c is the carrier-phase measurement error as well as the phase of the composite received signal.

A first-order model of the correlation function is given by

$$R(\tau) = \begin{cases} 1 - \frac{|\tau|}{T_c}, & |\tau| \leq T_c \\ 0, & |\tau| > T_c \end{cases}, \quad (15.10)$$

where T_c is the period of the GNSS code chip. This model ignores the influence of noise, assumes an infinite bandwidth signal and assumes the correlation function has no sidelobes (i. e., nonzero values outside of the main peak). If noise is ignored in this manner, then the

integration/averaging operations in (15.7)–(15.9) can be dropped.

An alternate form of (15.9) can be obtained by first considering the phasor diagram of Fig. 15.7 depicting the direct signal and a single multipath signal [15.8]. The relative phase of the multipath is θ_m whereas the phase of the composite received signal is θ_c . The phasors are given by

$$\begin{aligned} D &= a_0 R(\tau_c), \\ M &= a_1 R(\tau_c - \tau_1) e^{j\theta_m}, \end{aligned} \quad (15.11)$$

where, as before, τ_c is DLL code tracking error (depicted, for example, in the *lower right* of Fig. 15.6). To aid in the determination of θ_c , the multipath phasor is decomposed into in-phase (M_I) and quadrature (M_Q) components depicted in Fig. 15.8 [15.8]

$$\begin{aligned} M_I &= a_1 R(\tau_c - \tau_1) \cos(\theta_m), \\ M_Q &= a_1 R(\tau_c - \tau_1) \sin(\theta_m). \end{aligned} \quad (15.12)$$

Determination of θ_c is then a simple application of trigonometry

$$\begin{aligned} \theta_c &= \arctan \left(\frac{M_Q}{D + M_I} \right) \\ &= \arctan \left[\frac{\alpha_1 R(\tau_c - \tau_1) \sin(\theta_m)}{R(\tau_c) + \alpha_1 R(\tau_c - \tau_1) \cos(\theta_m)} \right], \end{aligned} \quad (15.13)$$

where $\alpha_1 = a_1/a_0$ is the relative amplitude of the multipath (also known as the multipath-to-direct ratio or M/D). For the general case of N multipath signals, the composite phase is given by [15.9]

$$\theta_c = \arctan \left(\frac{\sum_{i=1}^N \alpha_i R(\tau_c - \tau_i) \sin(\theta_{m,i})}{R(\tau_c) + \sum_{i=1}^N \alpha_i R(\tau_c - \tau_i) \cos(\theta_{m,i})} \right). \quad (15.14)$$

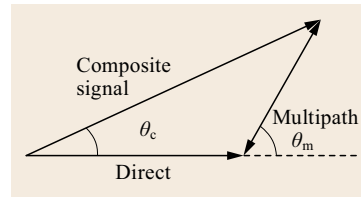


Fig. 15.7 Phasor diagram depicting the composite phasor as the vector sum of the direct and multipath phasors

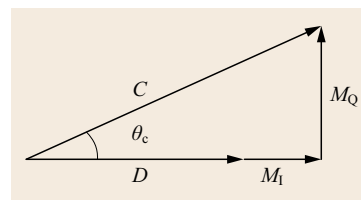


Fig. 15.8 Phasor diagram with the multipath phasor resolved into in-phase and quadrature components

Note that since θ_c spans 0 to 2π , a four-quadrant arctangent function must be utilized when performing computer simulations.

For simple reflections from planar surfaces, the relative delay can be calculated as discussed earlier. The

relative amplitude is determined as a combination of the reflection coefficient and any attenuation provided by surface roughness. For reflecting surfaces that do not fully span the first Fresnel zone, an additional attenuation factor applies as well.

15.4 Pseudorange and Carrier-Phase Error

Calculation of pseudorange and carrier-phase multipath error for the case of coherent DLLs is somewhat complicated due to the coupling of the equations. For example, (15.7) provides the equation for the discriminator function that is distorted due to multipath. The τ for which this function is zero corresponds to the pseudorange error due to the multipath (noted herein by τ_c). However, in addition to the parameters that specify the receiver's correlator spacing (d) and the multipath (a_i , θ_i , τ_i), the composite phase (θ_c) must also be specified. The composite phase is given in (15.14) but note this equation requires the pseudorange tracking error (τ_c) which, as just stated, is the solution of (15.7) when

it is set equal to zero. The simultaneous solution of these equations must thus be performed iteratively. Initial guesses for τ_c are used to calculate the discriminator value and slope and the initial guesses are also used in the calculation of the composite phase. Specifically, for a given value of τ used to evaluate the discriminator, that same value is first used to determine the composite phase. As the iteration continues in the search for the zero of the discriminator function, both τ_c and θ_c converge to the correct solution simultaneously. Of course, the situation is simpler in the case of noncoherent DLLs since the discriminator function (15.8) is not a function of the composite phase.

15.5 Multipath Error Envelopes

The pseudorange multipath error concept just illustrated can be quantified analytically and through simulation. Figure 15.9 provides an example of a BPSK pseudorange multipath error *envelope* for the case of a single multipath signal that is half the amplitude of

the direct signal (the relative amplitude is also referred to as the multipath-to-direct or M/D ratio).

Three cases are illustrated (all with the simplifying assumption of infinite signal bandwidth):

1. BPSK(1) tracking with 1 chip early-to-late spacing
2. BPSK(1) tracking with 0.1 chip spacing
3. BPSK(10) tracking with 1 chip spacing.

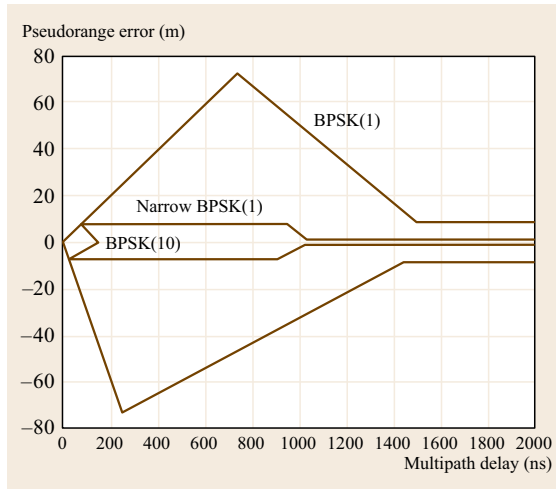


Fig. 15.9 Pseudorange multipath error envelopes for BPSK(1) wide correlator, BPSK(1) narrow correlator, and BPSK(10) wide correlator (infinite bandwidth assumed)

In each case, the upper curve represents the maximum error due to multipath and the lower curve represents the minimum error. For pseudorange measurements, the maximum error occurs when the relative phase of the multipath is zero (also referred to as *in-phase* multipath) and the minimum error occurs when the relative phase is 180° (also known as *out-of-phase*). The pseudorange error for all other multipath relative phases will lie somewhere between the upper and lower curves, hence the use of the term *envelope*. A general form of the infinite-bandwidth multipath error envelope (for BPSK signals) was derived in [15.10] and is illustrated in Fig. 15.10.

Several important points may be noted from this figure. First, the maximum pseudorange error is $\alpha d/2$ where α is the M/D ratio and d is the early-to-late correlator spacing. This provides an easily calculated upper bound on pseudorange multipath error for a sin-

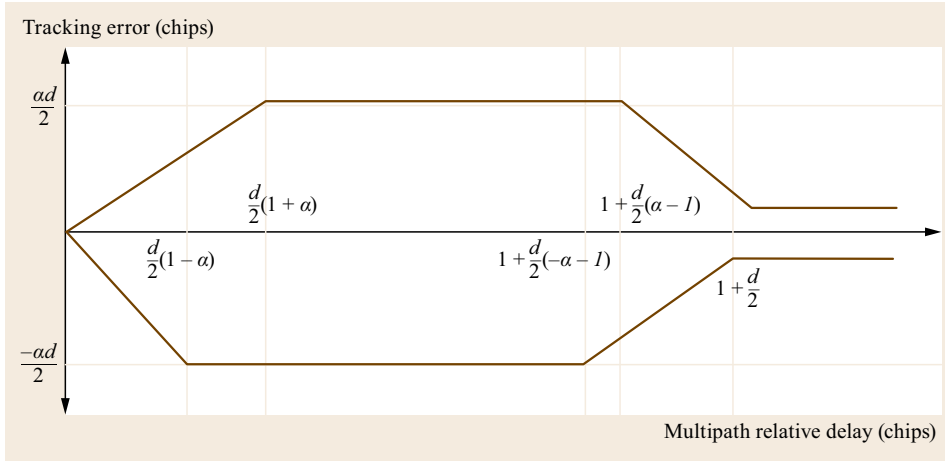


Fig. 15.10
Generalized
infinite-
bandwidth
multipath error
envelope for
BPSK tracking

gle specular reflection. Second, the maximum negative error occurs at a shorter relative delay than the maximum positive error and this is related to the fact that the error has a nonzero mean, especially when the M/D ratio is large. Third, the envelope does not decrease to zero at large relative delays. This is due to the presence of sidelobes in the autocorrelation function of the GNSS signal [15.10]. It should also be noted that the pseudorange multipath error envelopes are the same for both coherent and noncoherent DLLs. Behavior within the envelopes, however, are different and this manifests in differing average error characteristics as will be shown in a later section.

As described in Chaps. 7 and 9, newer signal structures have been developed based on binary offset carrier (BOC)/multiplexed binary offset carrier (MBOC)/composite binary offset carrier (CBOC) techniques. As shown in Chap. 4, the correlation functions of these signals have the same envelope as those of BPSK signals (with the same chipping rate) but differ via the sawtooth shape within the envelope. As a result, the general BPSK multipath error envelope (Fig. 15.10) bounds the corresponding envelopes of the BOC/MBOC/CBOC signals. An example is shown in Fig. 15.11 for $\alpha = 0.5$ and $d = 0.1$ (except for the BPSK(10) signal in which $d = 1$).

Calculation of the pseudorange multipath error envelope is simplified somewhat due to the decoupling of the pseudorange and carrier-phase errors. Specifically, maximum pseudorange error occurs when there is maximum distortion of the modulation of the received signal. As described above, this occurs under conditions of total constructive interference (i.e., the multipath is in-phase with the direct) or total destructive interference (i.e., out-of-phase multipath). Thus, the multipath relative phase is simply 0 or π . Such a simple relationship is not the case, however, when one attempts to calculate

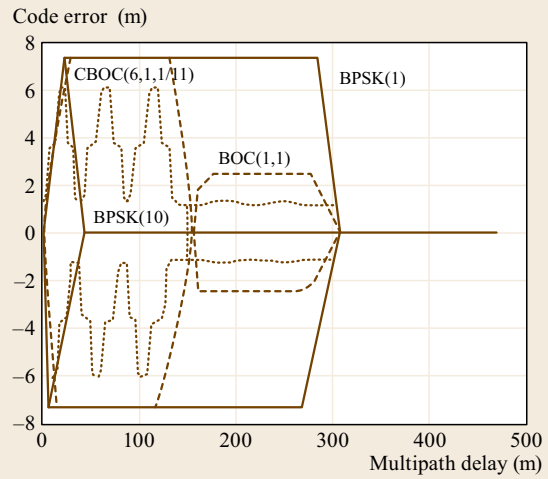


Fig. 15.11 Infinite-bandwidth multipath error envelopes for narrow correlator ($d = 0.1$ chip) for BPSK(1), BOC(1,1), CBOC(6,1,1/11) and standard correlator ($d = 1$ chip) for BPSK(10). The relative multipath amplitude is 0.5

the carrier-phase multipath error envelope. The theoretically exact result must be obtained through a search process. Specifically, for a given multipath relative delay (τ_m), the carrier-phase error equation (e.g., (15.13) and either (15.7) or (15.8)) must be solved for a range of relative phases between 0 and π . The envelope value is then given by the maximum value of θ_c obtained over all relative phases. This procedure was utilized in [15.8] and the results were validated with a real receiver and a hardware simulator.

The aforementioned procedure is cumbersome and simplified models have been described [15.7, 11]. As can be readily deduced from Fig. 15.7, the maximum carrier-phase error occurs when the multipath phasor is

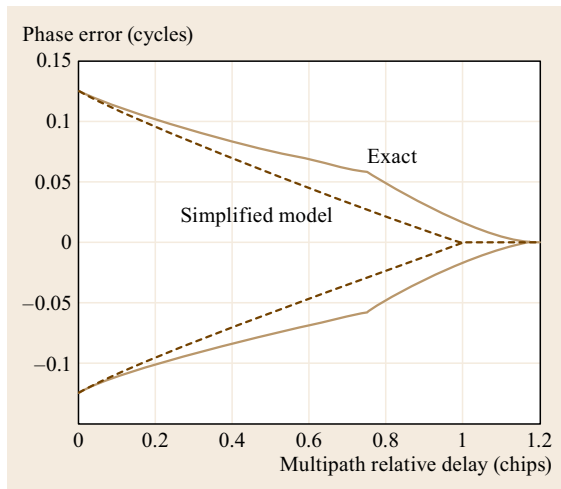


Fig. 15.12 Carrier-phase multipath error envelope (strong multipath, wide correlator)

orthogonal to the composite phasor. In this special case, the composite carrier-phase is given by [15.12]

$$\begin{aligned} \max(\theta_c) &= \arcsin(M/D) \\ &= \arcsin\left(\frac{\alpha_1 R(\tau_c - \tau_1)}{R(\tau_c) + \alpha_1 R(\tau_c - \tau_1)}\right). \end{aligned} \quad (15.15)$$

However, this equation is difficult to evaluate for coherent DLLs since, as discussed above, τ_c can only be determined after θ_m has been specified. *Van Nee* has noted [15.7] that a simplification can be achieved if the pseudorange tracking error (τ_c) is negligibly small. Specifically, if $\tau_c \approx 0$, then the first-order model of the correlation function (15.10) yields: $R(0) = 1$. Substituting these values into (15.15) yields the approximate formula [15.7, 8, 11]

$$\max(\theta_c) \approx \arcsin(\alpha_1 R(\tau_1)). \quad (15.16)$$

Figure 15.12 depicts the carrier-phase multipath error envelope for a coherent DLL with an early-to-late correlator spacing of 1 chip and an M/D of -3 dB. Both the theoretically exact result and the result with the approximate formula are shown. In this somewhat extreme case, there is a noticeable discrepancy between the two. However, even with such a strong multipath signal, if a narrow-correlator architecture (0.1 chip early-to-late spacing) is utilized, the approximation is

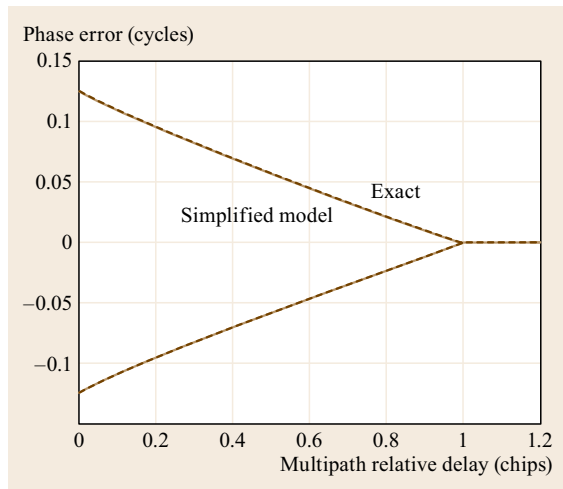


Fig. 15.13 Carrier-phase multipath error envelope (strong multipath, narrow correlator)

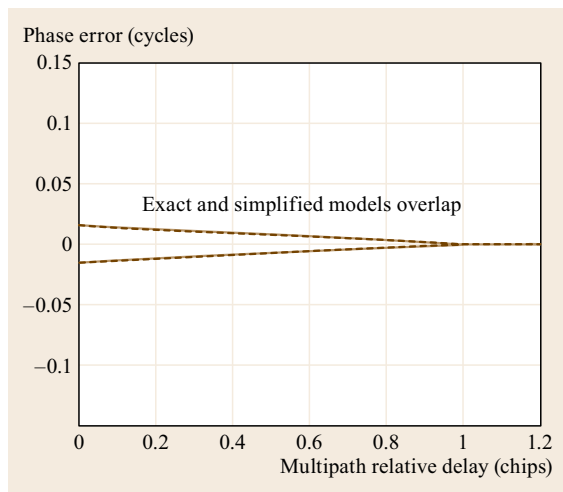


Fig. 15.14 Carrier-phase multipath error envelope (weak multipath, wide correlator)

quite good as shown in Fig. 15.13. Furthermore, for the case of weak multipath signals, the approximate formula is quite accurate even for wide correlator spacing (Fig. 15.14). Observe that, unlike the pseudorange, the carrier-phase multipath error envelope has its maximum value at a relative delay of zero. Thus carrier-phase multipath error increases as the receiver-antenna gets closer to a reflecting object.

15.6 Temporal Error Variation, Bias Characteristics and Fast Fading Considerations

Typically the multipath relative delay will vary with satellite/receiver/reflecting-obstacle motion. For example, in the case of a stationary receiver experiencing ground multipath (Fig. 15.3), the relative delay will start near zero as the satellite rises above the horizon, will increase to a maximum value of $2h$ when the satellite is directly overhead, and then will decrease back to zero as the satellite sets (the maximum value will be less if the satellite does not pass directly overhead).

To understand the temporal variation of the pseudorange and carrier-phase multipath error as relative path delay varies, Figs. 15.15 and 15.16 plot, for BPSK(1) and $d = 1$, the errors for a short range of relative delay with the relative multipath phase computed only as a function of the path delay. In these plots, the relative phase of the multipath is computed from path delay by converting the path delay to units of wavelengths, subtracting off the whole number of wavelengths, and then converting the remaining residual to radians (note this computation ignores any phase shifts that occur at the surface of the reflecting obstacle). From these figures, one can observe that the errors are largely sinusoidal for weak multipath relative amplitudes but are nonsinusoidal for strong multipath relative amplitudes. The errors also appear to be orthogonal. That is, the pseudorange errors appear to peak when the carrier-phase error is zero and vice versa.

This is not precisely correct. As discussed earlier, the multipath phasor must be orthogonal to the composite phasor to achieve maximum carrier-phase error. For strong multipath this will occur for relative phase values significantly greater than 90° . For weak multipath scenarios and/or multipath-mitigating receiver architectures such as narrow-correlator, the pseudorange, and carrier-phase errors are approximately orthogonal.

Furthermore, pseudorange multipath errors have nonzero mean values. This can be shown by averaging the pseudorange multipath errors, at a given value of relative delay, for relative phase values over the range of 0 to 2π . Figure 15.17 depicts the BPSK bias error for a coherent wide correlator spacing (1.0 chip) receiver with multipath relative amplitudes ranging from -20 to -3 dB. The bias error is most pronounced for strong multipath and relative delays in the range of one-half to one chip. As Fig. 15.18 shows, there is approximately a factor of 5 reduction in peak bias error with narrow correlator spacing (0.1 chip). However, for short delay multipath, the performance advantage disappears as shown in Fig. 15.19.

Similar characteristics may be observed with noncoherent receivers. Figure 15.20 depicts bias error for a noncoherent wide correlator spacing (1.0 chip) receiver with multipath relative amplitudes ranging from -20 to -3 dB. Although the peak error is nearly dou-

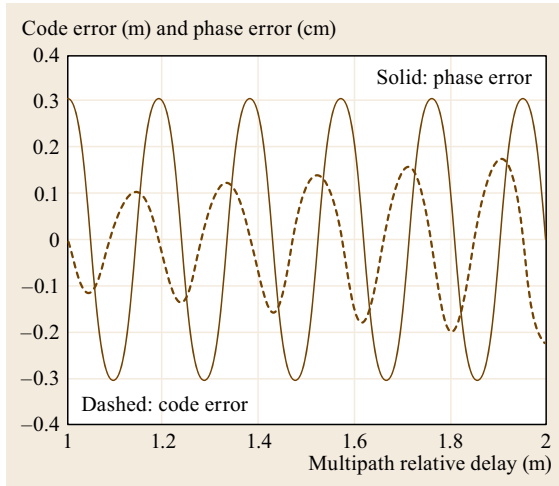


Fig. 15.15 Code and carrier-phase multipath error variation with multipath relative phase computed as a direct function of relative delay. The signal is BPSK(1) and the early-to-late correlator spacing is 1 chip (weak multipath case: $M/D = -20$ dB)

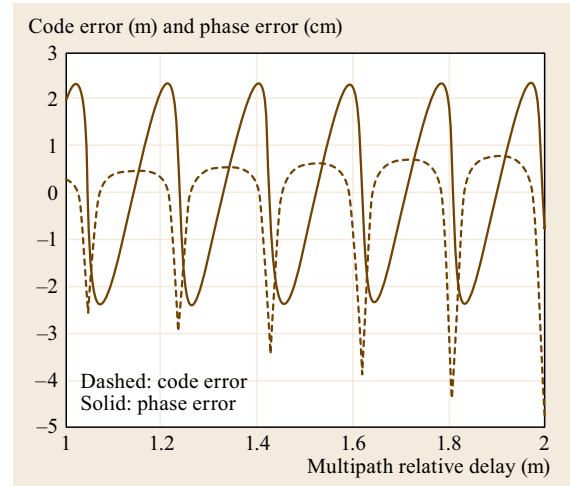


Fig. 15.16 Code and carrier-phase multipath error variation with multipath relative phase computed as a direct function of relative delay. The signal is BPSK(1) and the early-to-late correlator spacing is 1 chip (strong multipath case: $M/D = -3$ dB)

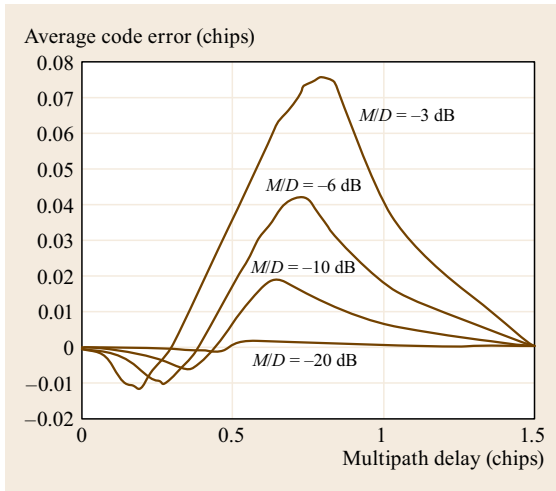


Fig. 15.17 Average pseudorange multipath error for a coherent BPSK receiver with 1 chip correlator spacing

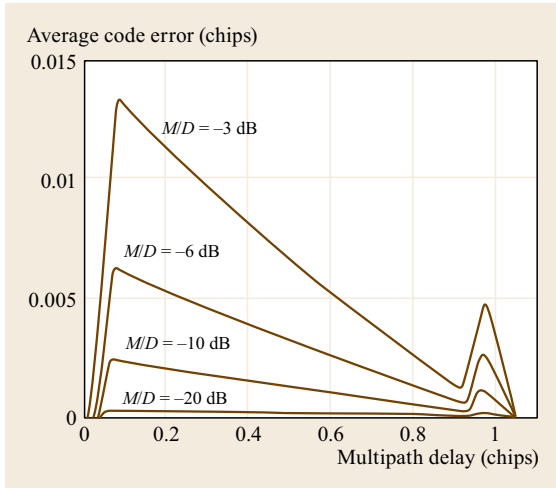


Fig. 15.18 Average pseudorange multipath error for a coherent BPSK receiver with 0.1 chip correlator spacing

ble the corresponding value for the coherent receiver, it may also be noted that the error is effectively zero for very short relative delays. The average error for noncoherent narrow-correlator receivers is depicted in Fig. 15.21. Although it is not immediately apparent, the average error for the noncoherent narrow correlator is slightly larger than the corresponding average error for the coherent narrow correlator receiver shown earlier. Again, the distinct difference in behavior of the wide and narrow-correlator receivers for short delay multipath may be observed in Fig. 15.22.

Thus, it is important to understand that simple averaging of measurements from a static receiver is not guaranteed to eliminate the effects of pseudor-

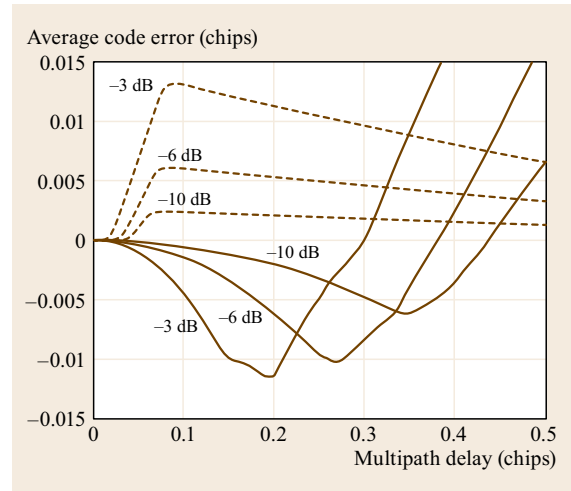


Fig. 15.19 Comparison of average pseudorange multipath error for wide correlator (solid lines) and narrow correlator (dashed lines) coherent BPSK receivers for short multipath relative delay

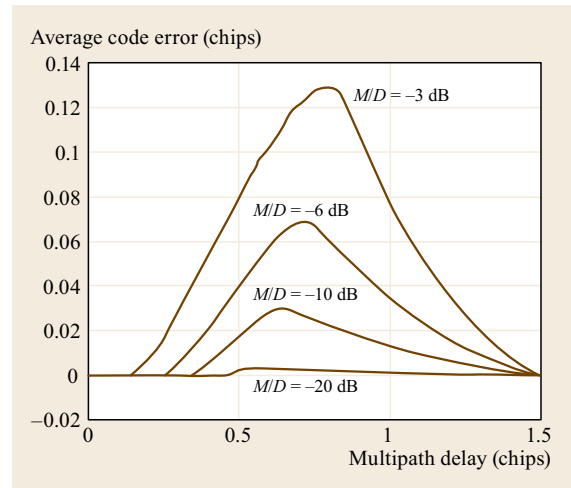


Fig. 15.20 Average pseudorange multipath error for a non-coherent BPSK receiver with 1 chip correlator spacing

ange multipath. However, the same is not true for the carrier-phase. Carrier-phase multipath error is in fact zero-mean and thus the error can effectively be reduced through averaging for static receivers.

As described earlier, dynamic receivers can induce relatively high phase rates in the received multipath signal. Detailed simulation and laboratory experiments have shown that typical GNSS receiver tracking loops are relatively insensitive to multipath signals with fading frequencies that exceed the loop bandwidth. Since carrier tracking loops are typically quite wide (e.g., 10–25 Hz), this phenomenon is not of much practi-

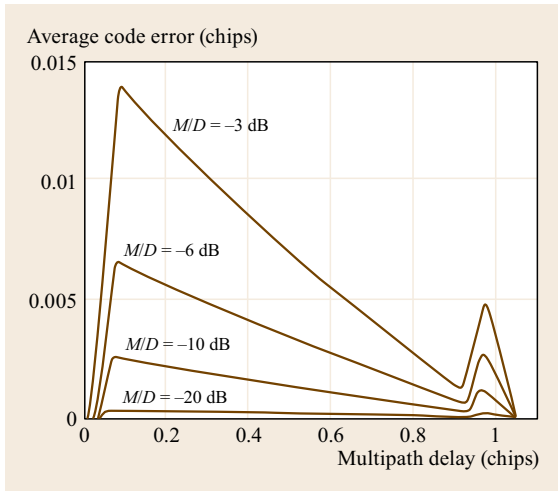


Fig. 15.21 Average pseudorange multipath error for a noncoherent BPSK receiver with 0.1 chip correlator spacing

cal use for carrier-phase multipath reduction. However, many receivers are constructed with carrier-aided code tracking loops and, in such receivers, the code tracking loop can be quite narrow (e.g., 0.05 Hz). Significant pseudorange multipath error reduction can be achieved for situations in which receiver dynamics induce multipath fading frequencies well above 0.1 Hz [15.8, 13].

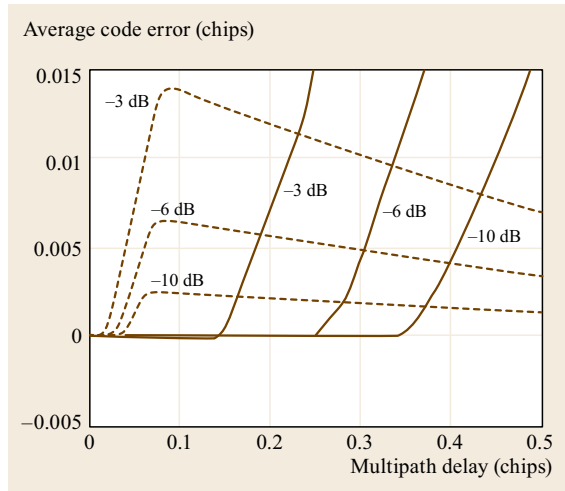


Fig. 15.22 Comparison of average pseudorange multipath error for wide correlator (*solid lines*) and narrow correlator (*dashed lines*) noncoherent BPSK receivers for short multipath relative delay

One must be careful, however, not to assume that the error is attenuated linearly with multipath fading frequency. It has been shown that a bias can remain that is a function both of the fading frequency and the multipath relative delay [15.13].

15.7 Multipath Mitigation

There are four broad categories of multipath mitigation techniques:

- Antenna placement
- Antenna type
- Receiver type
- Measurement post-processing.

15.7.1 Multipath Mitigation via Antenna Placement

Although it is a statement of the obvious, the best way to mitigate multipath is to place the antenna in a low or ideally multipath-free environment. The location for a permanent differential ground reference station, for example, should be chosen only after careful consideration of the multipath environment. Frequently, in such cases, there are conflicting requirements. For example, the most ideal, low-multipath environment is a flat, open field. The sole source of multipath, ground reflection, could be eliminated by placing the antenna flush on the ground itself. However, this is not typically practical since the antenna could be inadvertently cov-

ered by snow, ice or leaves and could be damaged by ground-keeping equipment. Thus, even in open fields, differential ground reference station antennas are typically mounted on pedestals and some ground reflection is accepted.

In practice, it is more common to encounter a non-ideal environment and still need to minimize the impact of multipath. Small changes in antenna location can still have a significant impact. This may be highlighted with a specific situation encountered by the author of this chapter. A differential ground reference station had been installed on top of the elevator tower of a multi-story parking garage (Fig. 15.23). After many weeks of operation, it was determined that multipath was affecting the performance but was doing so inconsistently. Specifically, there were long periods of time in which the multipath contamination was negligible and yet there were also long periods of time in which the multipath was highly problematic.

Using multipath measurement techniques (that will be described later in this chapter), it was eventually determined that when the top level of the garage was empty, the antenna was subject to specular reflection

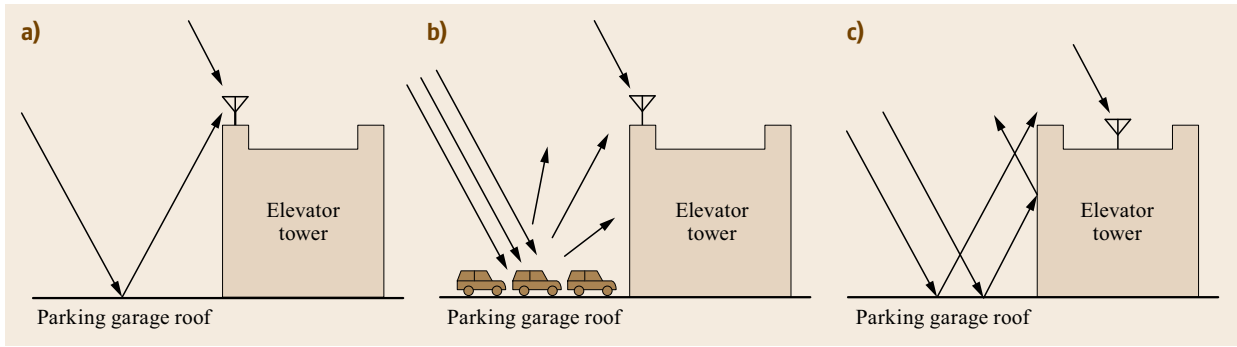


Fig. 15.23a–c Ground reference antenna siting case study: When there were no cars on the top level of the parking garage, there were strong specular reflections in the direction of the differential ground reference station antenna (a). The presence of cars on the top level of the parking garage induced diffuse multipath with relatively little energy directed toward the differential ground reference station antenna (b). By placing the antenna in the center of the elevator tower roof and at the level of the surrounding wall, the reflections from the top level of the garage were blocked by the elevator tower itself (c)

as shown in Fig. 15.23a. However, when the top level was full of cars, the reflections became diffuse thus reducing the strength of the multipath impinging on the antenna. The solution was to move the antenna to the middle of the elevator tower roof and let the tower block the reflections. The elevator tower roof was surrounded by a 1 m high wall and the antenna was installed at a height that was even with the top of this wall. The height of the antenna above the roof itself prevented the antenna from being covered with snow during the winter months. This installation worked extremely well by reducing the multipath to negligible levels.

15.7.2 Antenna Type

With a given environment, the next line of defense against multipath is to choose an antenna that attenuates the multipath while preserving the strength of the (desired) direct signal. In differential ground reference stations, for example, the primary source of multipath is the ground itself. The multipath is thus incident from negative elevation angles whereas the desired direct signal is incident from positive elevation angles. An ideal ground reference station antenna would have uniform gain above a given mask angle and zero gain below it. Two broad categories of antennas have been developed in an attempt to approximate this ideal: (a) single-element antennas mounted on specialized groundplanes; (b) fixed-beam phased-array antennas.

In the mid 1980s, research was conducted on the use of radio frequency (RF) absorber material to reduce the effect of multipath. A single-element global positioning system (GPS) antenna was mounted on a planar slab of RF absorber. Test results indicated at least a 30% reduction in pseudorange error due ground reflections could

be achieved by using the absorber material [15.14]. Despite the promising results, however, the technique was not widely adopted by the GNSS community. This was due in part to the expense of the material and challenges associated with maintaining its integrity under various weather conditions. Furthermore, the development of the choke ring antenna proved to be superior.

The choke ring design was originally developed by the Jet Propulsion Laboratory and the University of New Brunswick [15.15] shortly after the aforementioned RF absorber studies. The design consists of a single-element antenna mounted inside a set of concentric conductive rings. The rings effectively present a capacitive series reactance to the incident electric field and thus *choke out* electromagnetic waves arriving at low and negative elevation angles [15.15]. The antenna system thus has low gain at low and negative elevation angles. Given the relative ease of manufacture, rugged design and relative compactness, the choke ring antenna became the standard multipath-reducing antenna of choice for ground reference stations, particularly for surveying applications.

Although the choke ring design was, and continues to be, very popular, it is not considered portable since the outer diameter of the ring system is approximately 36 cm and the overall antenna system typically weighs over 4 kg. In 2000, NovAtel introduced a proprietary design to provide choke ring performance with significantly reduced size and weight [15.16, 17]. The design, known as a *pinwheel* antenna, consists of an array of 12, aperture coupled, spiral slots surrounded by 11 concentric slot rings. The antenna yields nearly identical performance to a choke ring design but is half the diameter and one-eighth the weight.

Although the choke ring and pinwheel antennas work well for surveying applications, both are ham-

pered by low gain between 5 and 15° elevation. Both types of antennas have gain patterns that peak at zenith (90° elevation) and gradually roll-off as the elevation angle decreases (with the lowest gain occurring for an elevation angle of -90°). The benefit is very low gain at negative elevation angles and thus ground reflections are suppressed. The penalty, however, is that the direct LOS signal is severely attenuated for elevation angles less than about 15° and thus both pseudorange and carrier-phase measurements will be very noisy. This gradual roll off in the antenna pattern is an unavoidable result of the antennas being electrically *small*. That is, both antenna types are small in terms of L1 wavelengths. The choke ring has a diameter of approximately 2 wavelengths and the pinwheel is approximately 1 wavelength in diameter. For most surveying applications, mask angles are typically 15° (primarily to avoid tropospheric spatial decorrelation) and thus the poor performance of these antennas below that angle is not of concern.

There are applications, however, in which mask angles must be set at 5° or even slightly lower in order to provide sufficient system availability. A prime example of this is the ground-based augmentation system (GBAS). GBAS is a form of differential GNSS that provides guidance information for aircraft landing in low visibility conditions (Chap. 31). Without the use of satellites between 5 and 15° elevation (as well as those that are higher), the availability of the system (i.e., the percentage of the day in which the system is operational) would not be high enough to be of practical use.

For an antenna to provide sufficient gain at the 5° mask angle and simultaneously attenuate signals at negative elevation angles, an electrically large phased array is needed. This basic concept was first described in 1994 [15.18, 19] and two basic design variations were contemplated. The first to be developed and field-tested [15.20] consisted of a vertically stacked linearly polarized dipole array along with a single circularly polarized antenna element mounted in a concave reflector [15.21]. The 16-element dipole array provided coverage from 5 up to 35° elevation. The mini dish antenna provided coverage from 30° up to zenith. Separate GPS receivers, tied to a common clock, were fed by the two antenna subsystems and a downstream central processing unit (CPU) ran software that dealt with the hand-off of satellites between antennas for rising or setting satellites.

The second of the two design variants consisted exclusively of a vertical stacked array of circularly polarized elements [15.22]. The primary advantage of such an antenna is the need for only a single receiver and no complex downstream software to accommodate

satellite hand-off between antenna subsystems. Significant design challenges extended the development period [15.23] but it eventually became the standard ground reference station antenna for the United States version of GBAS.

Before closing this section it should be noted that promising techniques in novel ground plane design were introduced in 2010. The performance of one of these designs is highlighted in the carrier-phase multipath measurement section later in this chapter.

15.7.3 Receiver Type

As shown in Fig. 15.9, there is a distinct difference in pseudorange multipath error for wide and narrow correlator receiver architectures. Further, the size of the multipath error envelope scales inversely with chipping rate (e.g., BPSK(1) versus BPSK(10)). The narrow correlator principle can be applied to higher rate codes but the performance improvement is not as significant due to bandwidth limitations on the broadcast signal.

Starting in the mid-1990s, however, receiver designs were developed to exploit the full bandwidth of the slower chipping rate signals. In the case of GPS, these designs processed the coarse/acquisition (C/A)-code signal yet achieved nearly P(Y)-code multipath performance. Some examples include the Strobe correlator [15.24], enhanced Strobe correlator [15.25], designs by Leica [15.26], and so-called *superresolution* [15.27, 28]. The general concepts were described along with a theoretical noise analysis in [15.29]. One of two primary techniques were utilized. Gating or windowing of the received signal around the chip edges yielded effectively narrower chips and thus commensurately scaled down multipath error envelopes [15.25, 29].

The second technique combines two early and two late correlators in order to create a very narrow discriminator function. Improved multipath performance is thus achieved since the size and shape of the pseudorange multipath error envelope is proportional to the size and shape of the discriminator function. The four correlators associated with this technique are depicted in Fig. 15.24 for the case of a BPSK correlation function. Note the first pair (E_1, L_1) are separated by d chips and the second pair (E_2, L_2) are separated by $2d$. The discriminator function is constructed as the difference of two narrow-correlator discriminator functions [15.29, 30]

$$\begin{aligned} D_{\Delta\Delta} &= (E_1 - L_1) - \frac{1}{2}(E_2 - L_2) \\ &= D_{\text{narrow}}(d) - \frac{1}{2}D_{\text{narrow}}(2d) . \end{aligned} \quad (15.17)$$

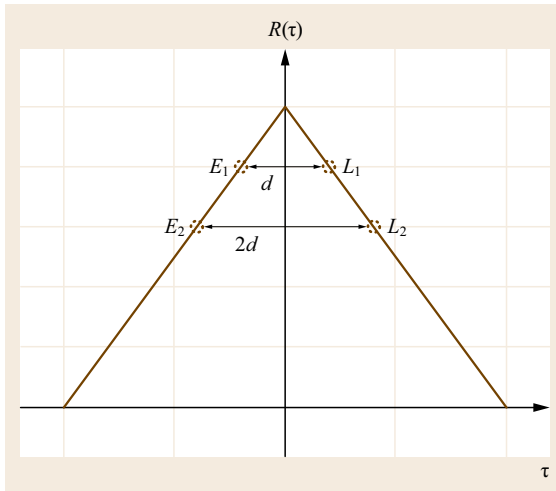


Fig. 15.24 There are two sets of early/late correlators in the double delta. One set has a separation of d chips and the other has a separation of $2d$ chips

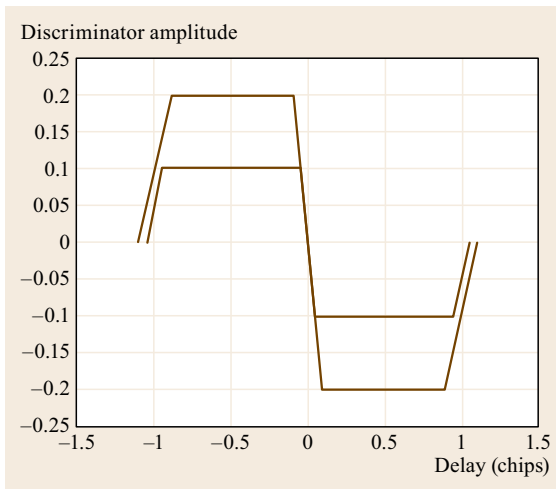


Fig. 15.25 The double delta can be viewed as a combination of two narrow correlator discriminator functions. This figure illustrates the two constituent functions for BPSK with $d = 0.1$ chip. The curve that peaks at ± 0.2 is the discriminator function for $2d = 0.2$ chip, the curve that peaks at ± 0.1 for $d = 0.1$ chip. These curves are combined using (15.17) to yield the double-delta discriminator shown in Fig. 15.26

Although this technique was implemented by manufacturers with names such as the *Strobe correlator* [15.25] and *pulse aperture correlator* [15.31], it was first generalized as the *high resolution correlator* [15.29]. However, the technique became most widely known as the *double-delta correlator* since it effectively consists of the difference of two differ-

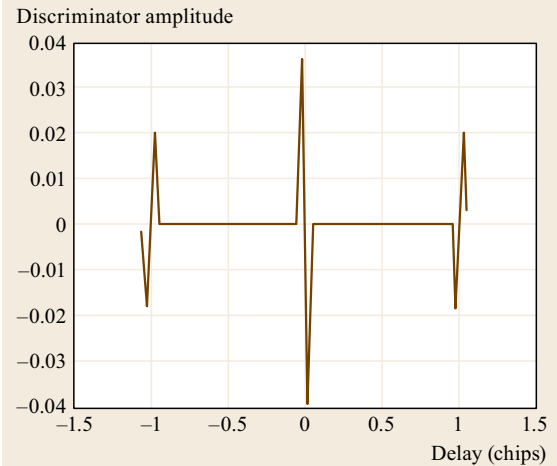


Fig. 15.26 The double-delta discriminator consists of a narrow *s-curve* region along with parasitic responses at plus and minus one chip delay (BPSK with $d = 0.1$ is depicted)

ences [15.30]. The two constituent narrow correlator discriminator functions are depicted in Fig. 15.25 and the resulting double-delta discriminator is depicted in Fig. 15.26 for the case of $d = 0.1$ chip (both for BPSK). The multipath envelopes for the double-delta constructed from 0.1 chip to 0.2 chip narrow correlators are depicted in Fig. 15.27 for the case of BPSK(1) where the relative multipath amplitude is -6 dB. It should be noted that in many environments, multipath is predominately short and medium delay and thus the echoes in the double-delta multipath error envelope at a delay of approximately 1 chip are not of serious concern.

Echoes notwithstanding, the double-delta correlator thus achieves a narrow discriminator function and narrow multipath envelope with the use of four correlators as opposed to the conventional single early and single late correlators used in a traditional DLL. Subsequent research has shown that almost arbitrary shaping of the discriminator function (also known as the *S-curve*) can be achieved with the use of additional correlators. This technique has been referred to as *S-curve shaping* and researchers have investigated the development of optimum *S-curves* given the use of a large number of correlators [15.32, 33].

15.7.4 Measurement Processing

This final category of mitigation involves dealing with multipath-contaminated measurements. There are numerous applications in which multipath is unavoidable despite the best attempts at optimum choice of antenna placement, antenna design, and receiver architecture.

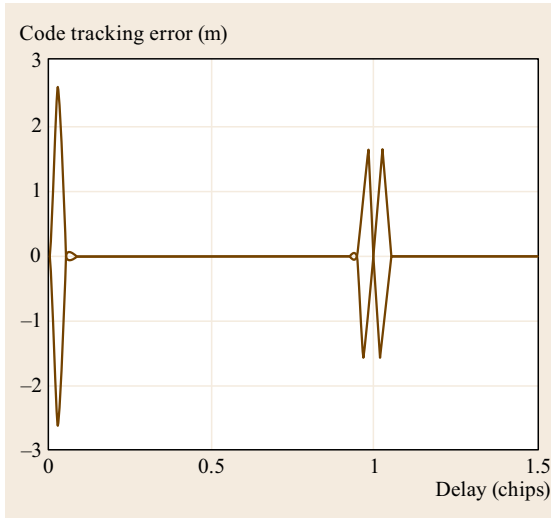


Fig. 15.27 The double-delta pseudorange multipath error envelope for BPSK(1) and $d = 0.1$ for the case where the multipath is 6 dB below the direct signal

The choice of measurement-processing mitigation technique depends on whether real-time operation is needed or not and whether the receiver is stationary or not.

The most challenging scenario, of course, is the need to mitigate multipath at the measurement level for a dynamic receiver in real time. The simplest technique involves elevation-angle dependent measurement weighting in a generalized least-squares solution or in a Kalman filter. The key idea is that, in most applications, multipath contamination of the measurements is approximately inversely proportional to the elevation angle of the satellite (note that this principle is true of tropospheric and ionospheric errors as well). The variances of the ranging measurements can be divided by the sine of the elevation angle to provide a simple, but reasonable approximation.

Beyond simple weighting schemes, real-time multipath mitigation in a dynamic receiver requires some technique to monitor the presence of multipath in the measurements. As will be described in the next section, it is possible to observe, in post-processing, the pres-

ence of multipath on pseudorange measurements via a judicious combination of the pseudorange measurement with the carrier-phase measurement. In real time, the difference between the raw pseudorange measurement and the output of a Hatch filter (carrier-smoothed pseudorange) can be computed as a metric of the combination of noise and multipath. The measurement can then be deweighted in proportion to the magnitude of the metric.

For non real-time applications, two techniques may be utilized for mitigation depending on whether the receiver is static or dynamic. If the receiver is dynamic, then the pseudorange multipath measurement technique, described in the next section, may be used to determine the magnitude of multipath contamination and then used to weight the measurements accordingly. If the receiver and its surroundings are static, then the multipath error will repeat daily in conjunction with the repeat of the satellite ground track. This phenomenon of multipath repeatability will be explored more in the next section but it can be exploited to reduce the impact of multipath at fixed sites (such as crustal monitoring and structural deformation monitoring applications). The technique is known as *sidereal filtering* since the GPS satellite ground tracks repeat approximately once every sidereal day. Put simply, the position solution results over the course of a day are low-pass filtered to estimate the multipath error and this correction is then removed from the results on subsequent days in order to observe small changes. Detailed descriptions may be found in [15.34, 35].

Specialized hardware–software have also been developed to monitor for abnormal multipath error (along with other signal deformations). The advent of so-called software defined receiver architectures, or more generally, receivers with more than three correlators per channel (in some cases, thousands of software correlators per channel), have enabled the development of signal quality or signal deformation monitoring. The use of many correlators per channel allows the correlation peak to be analyzed for deviations from nominal. The use of these techniques to detect multipath has been described, for example, in [15.36–39].

15.8 Multipath Measurement

It is frequently necessary to evaluate the impact of multipath using actual measurements. Measuring multipath presents challenges since there are a variety of GNSS error sources. In addition to multipath, pseudorange and carrier-phase measurements are impacted by satellite and receiver clock offsets, ionospheric and tro-

pospheric delays and receiver noise and tracking error. Thus a simple differencing of a measurement with the truth would yield the sum of all error sources and not any one in particular. What is needed is a technique to isolate the multipath error from the rest. The typical approach to assess multipath error on pseudorange

measurements is to exploit the commonality of a subset of the errors on both the pseudorange and carrier-phase. Isolation of carrier-phase multipath involves the use of carrier-phase double differencing of measurements obtained at a test site and at a reference site. Each of these two techniques will be described in turn.

15.8.1 Isolation of Pseudorange Multipath

Models of the pseudorange and carrier-phase measurements are given as follows

$$p = \rho + c dt_r - c dt^s + I + T + M_p + e + D_p, \quad (15.18)$$

$$\varphi = \rho + c dt_r - c dt^s - I + T + M_\varphi + \varepsilon + D_\varphi + N\lambda, \quad (15.19)$$

where all terms are in units of distance:

p	pseudorange measurement
φ	carrier-phase measurement
ρ	true satellite-to-receiver range
$c dt_r$	receiver clock offset
$c dt^s$	satellite clock offset
I	ionospheric delay
T	tropospheric delay
M_p	pseudorange multipath error
M_φ	carrier-phase multipath error
e	pseudorange noise
ε	carrier-phase noise
D_p	dynamics-induced code-tracking loop delay
D_φ	dynamics-induced carrier-tracking loop delay
$N\lambda$	carrier-phase range integer ambiguity.

The technique to isolate the pseudorange multipath error relies on the fact that the true range, satellite and receiver clock offsets, and the tropospheric delay are common both to the pseudorange and carrier-phase measurements [15.40–42]. By differencing the pseudorange and carrier-phase measurements for a given satellite, the true range, clock offsets and tropospheric delay all cancel

$$p - \varphi = 2I + (M_p - M_\varphi) + N\lambda + (D_p - D_\varphi) + (e - \varphi). \quad (15.20)$$

This observable (sometimes referred to as *code-minus-carrier* or **CMC**) has twice the ionospheric error since the ionospheric delay is equal in magnitude, but opposite in sign, for the pseudorange and carrier-phase measurements (the opposite effect of the ionosphere on the pseudorange and carrier-phase is referred to as *ionospheric divergence*). In most cases, the carrier-phase noise and multipath errors can be considered negligible, compared to those of the pseudorange, since they

are typically smaller by approximately 2 orders of magnitude. This simplifies the multipath observable to

$$p - \varphi \approx 2I + M_p + N\lambda + (D_p - D_\varphi) + e. \quad (15.21)$$

As would be expected, tracking loop delays due to dynamics are only a concern when a receiver is experiencing nontrivial acceleration. Obviously, this is not an issue for static receiver data collections. Furthermore, most receivers designed for vehicle applications have sufficiently wide tracking loop bandwidths to accommodate all normal vehicle maneuvers. In fact, even in receivers designed for static applications the carrier tracking loop bandwidth is wide enough to accommodate medium dynamics such as 2g turns in an aircraft. However, the code tracking loops in some receivers designed for static applications (e.g., surveying) are too narrow (and not carrier-aided) and exhibit lags during dynamics if the receiver is used in a vehicle. Sometimes this situation occurs when a survey-grade receiver is being used as a truth-reference system during a flight test. Vehicle dynamics result in nontrivial tracking error in the pseudorange but not in the carrier-phase. This difference shows up in the multipath observable and care must be taken not to confuse the effect with multipath error [15.42].

For static data collections, the dynamic terms are zero and the multipath observable reduces to

$$p - \varphi \approx 2I + M_p + N\lambda + e. \quad (15.22)$$

Assuming the carrier tracking loop has not experienced any cycle slips (a reasonable assumption if the carrier-to-noise ratio is strong and the receiver is static), the integer ambiguity is constant and thus shows up as a bias in the multipath observable. It can be removed simply by estimating and subtracting off the mean value of the multipath observable. This process also removes any bias in the multipath error component, however, so the result only depicts peak-to-peak behavior.

The ionospheric divergence can be handled in two ways. First, under normal conditions (e.g., in the absence of solar storms inducing ionospheric scintillation), ionospheric delay is strongly dependent upon the satellite elevation angle and thus exhibits itself as a long-term trend in the multipath observable. Roughly speaking, the ionospheric delay is inversely proportional to the satellite elevation angle. Since a satellite pass (across the sky for a static observer) has a duration of approximately 3–8 h, the ionospheric delay looks like a first- or second-order trend in a multipath observable that is less than a couple of hours in length. Thus the ionospheric term can be eliminated by fitting and removing a first- or second-order curve from

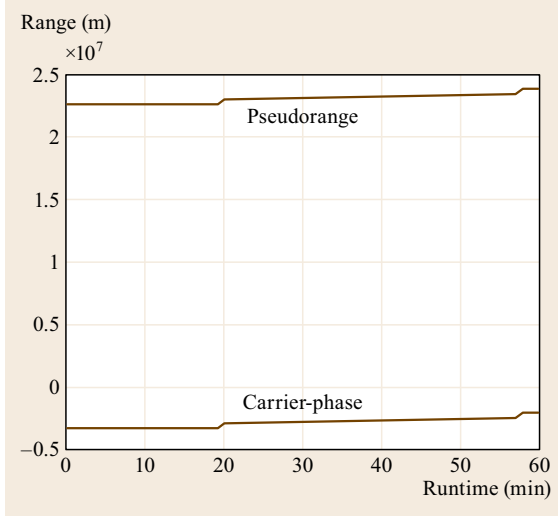


Fig. 15.28 Pseudorange and carrier-phase for pseudo-random noise (PRN) 20 GPS C/A-code. The data was collected from 1300 to 1400 h, GPS time, on day 350 of 2012 at the STKR continuously operating reference station (CORS) site in Athens, Ohio, USA

the multipath observable. A visual inspection of the raw observable is used to determine whether a first- or second-order curve-fit is appropriate.

The second technique can be applied if dual-frequency measurements are available. The ionospheric delay is estimated by forming the usual dual-frequency correction but doing so with carrier-phase measurements instead of pseudorange measurements. Since the carrier-phase noise and multipath are negligible compared to the pseudorange, the use of this correction does not severely impact the resulting observable. Of course, an ionospheric delay estimated by carrier-phase measurements will have a bias error (due to the range ambiguities) but, as discussed earlier, the multipath observable already has an unknown bias that has to be removed anyway. A closed-form expression for the dual-frequency iono-corrected code-minus-carrier observable is given by [15.43]

$$\text{CMC} = p_{L1} - \frac{f_1^2 + f_2^2}{f_1^2 + f_2^2} \phi_{L1} - \frac{2f_2^2}{f_1^2 + f_2^2} \phi_{L2}, \quad (15.23)$$

where f_1 and f_2 are the L1 and L2 carrier frequencies. There is nothing unique about these two frequencies, however, and the observable could be formed with measurements, for example, from L1 and L5. It should be noted the observable still has a carrier-phase ambiguity and thus the mean needs to be removed.

To illustrate this procedure, the results of a multipath analysis are presented for a rooftop CORS station

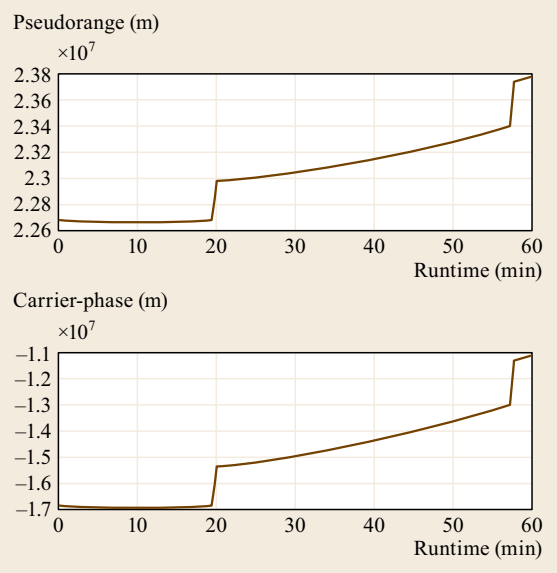


Fig. 15.29 Zoomed view of the pseudorange and carrier-phase for PRN 20 GPS C/A-code. Two clock jumps are clearly visible in both measurements

(STKR) located at Ohio University in Athens, Ohio, USA. The data was collected from 13:00 to 14:00 h (GPS time) on day 350 (December 15) of 2012. The pseudorange and carrier-phase measurements for GPS PRN 20 are shown in Fig. 15.28. Figure 15.29 shows a zoomed view that reveals two receiver clock jumps that occur simultaneously on both measurements and thus will cancel when the multipath observable is formed.

Figure 15.30 shows the results when the code-minus-carrier multipath observable is formed and the bias is removed. A long-term trend may be observed and is due to the ionospheric divergence. The medium frequency components are multipath and the high frequency components are noise.

Since the collected data included dual-frequency measurements, the ionospheric divergence can be removed by forming the dual-frequency ionospheric correction with the carrier-phase. The results are shown in Fig. 15.31. Note the long-term trend has been eliminated.

15.8.2 Short-Delay Multipath

The STKR CORS site consists of a roof-top mounted antenna. As a result, the dominant reflections originate from the roof itself and thus are *short* delay (i.e., less than 0.1 chip). As shown in the multipath error envelopes described earlier, both the BPSK(1) and BPSK(10) error envelopes overlap for short relative

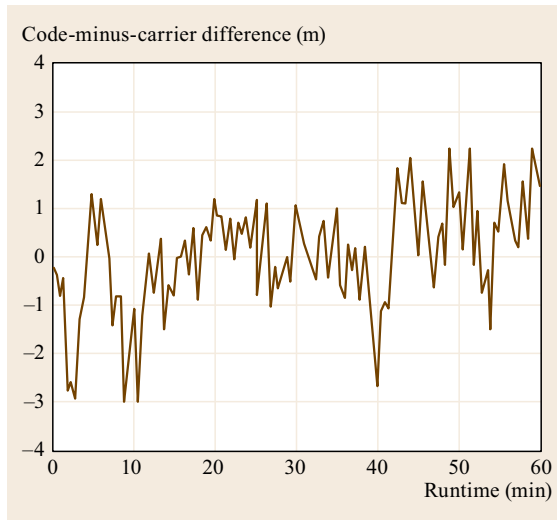


Fig. 15.30 Code-minus-carrier multipath observable for PRN 20 GPS C/A-code. The bias in the observable has been estimated and removed. The long-term trend in the data is the ionospheric divergence. The medium frequency behavior is pseudorange multipath and the high-frequency components are noise

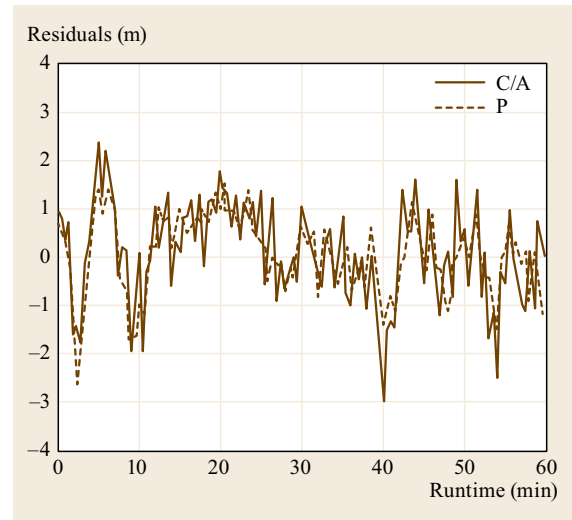


Fig. 15.32 Ionosphere-corrected code-minus-carrier multipath observable for PRN 20 GPS C/A-code and P-code. The low/medium frequency components are due to multipath and are largely the same on both measurements. This is due to the short relative delay of the multipath

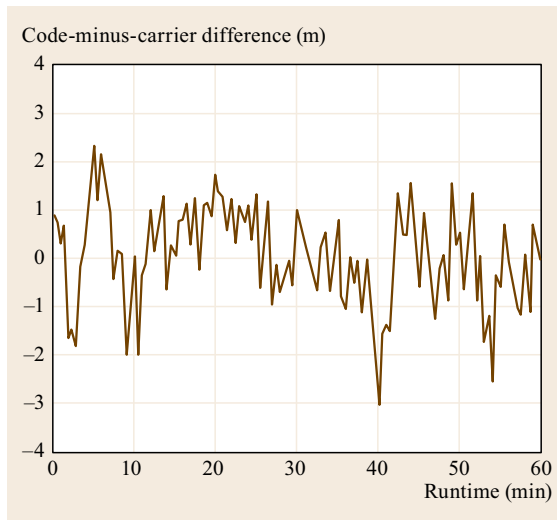


Fig. 15.31 Code-minus-carrier multipath observable for PRN 20 GPS C/A-code. The long-term trend due to ionospheric divergence has been eliminated through the application of a dual-frequency ionospheric correction based on carrier-phase measurements

multipath delay. This is illustrated in Fig. 15.32. Although the GPS P-code observable has less noise (as expected), both the P and C/A-code show similar multipath error characteristics.

15.8.3 Multipath Repeatability

When a multipath environment is static (stationary receiver and multipath-producing objects), satellite ground-track repeatability can be exploited to help distinguish multipath from noise excursions. In GPS, for example, the satellite ground track repeats after approximately 23 h and 56 min. In a static multipath environment, the multipath error will thus also repeat along with the satellite ground-track. Thus, if a certain multipath error occurs, say, on a Tuesday at 4:00 pm, then that same error will occur again on Wednesday at approximately 3:56 pm. If a particular characteristic is observed in the ionosphere-corrected code-minus-carrier observable on separate days, it may be concluded that the characteristic is due to multipath and not noise.

To illustrate this technique, the results of a repeatability analysis are presented in the next series of plots. In addition to the data collected on day 350 of 2012 (discussed earlier), data was also analyzed from day 345 and day 349. Figure 15.33 depicts the PRN 20 C/A residuals for each of the 3 days (the residuals from the first and third days are intentionally biased by +4 and -4 m, respectively, to provide visual clarity and the 4 min/d shift has also been taken into account). When the code-minus-carrier observables for all 3 days are plotted together (Fig. 15.34), the multipath error oscillation during the first 10 min of the data collection period is clearly visible.

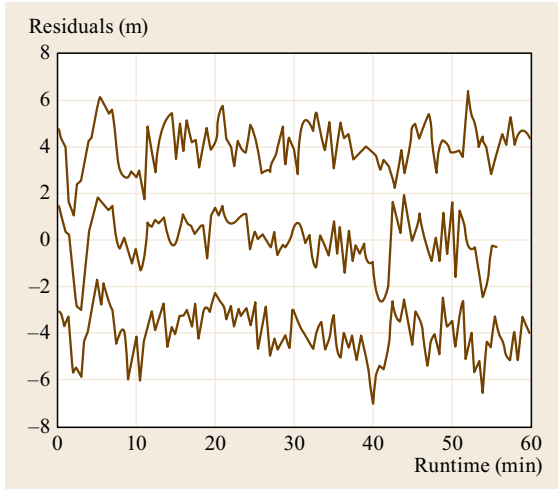


Fig. 15.33 Multipath repeatability analysis for PRN 20 GPS C/A-code. The data was collected from 1300 to 1400 h, GPS time, on days 345, 349, and 350 of 2012 at the STKR CORS site in Athens, Ohio, USA. The code-minus-carrier observables for day 345 are intentionally biased by +4 m and those of day 350 are biased by -4 m for visual clarity

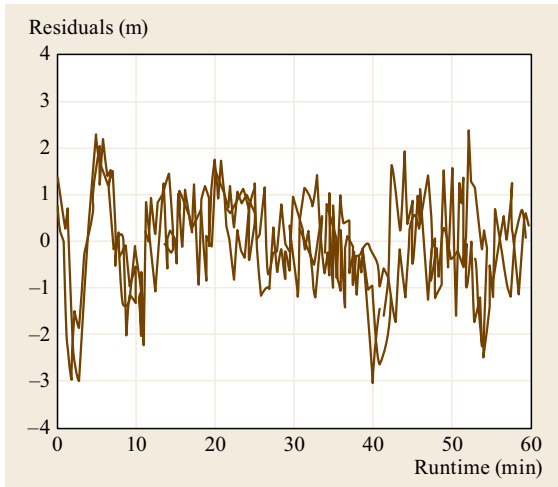


Fig. 15.34 The code-minus-carrier observables of Fig. 15.33 for all 3 days are plotted on top of each other. The multipath oscillation during the first 10 min is clearly visible in all three

The results for PRN 04 are depicted in Fig. 15.35. In this case the multipath error appears to be about at the same level as the noise. The slightly lower noise P-code results (Fig. 15.36) do not, at first, clearly distinguish between noise and multipath. When the results from the 3 days are overlayed, (Fig. 15.37), however the multipath errors become much more apparent.

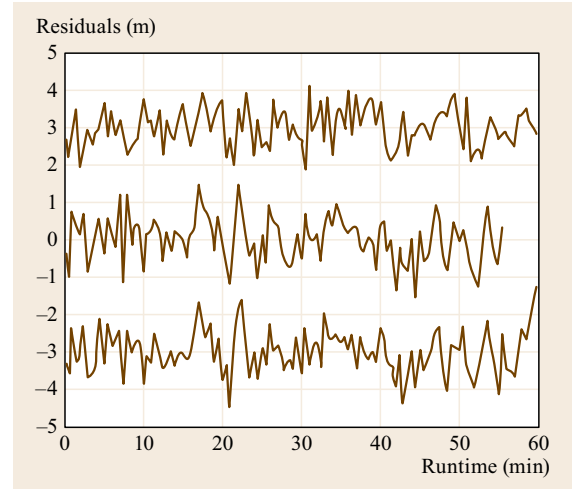


Fig. 15.35 Same as Fig. 15.33 for PRN 04 GPS C/A-code

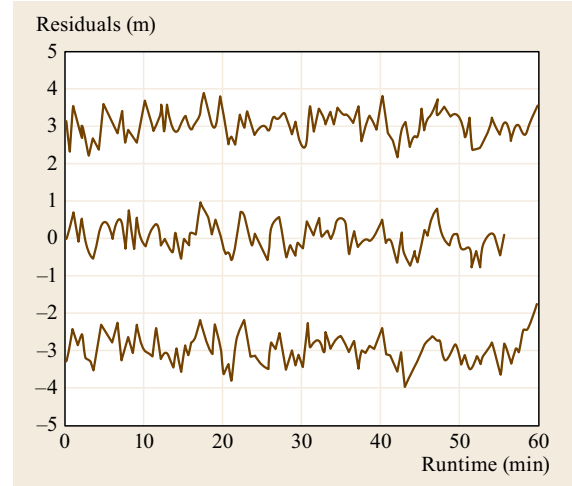


Fig. 15.36 Same as Fig. 15.33 for PRN 04 GPS P-code

The lower impact of multipath on PRN 04 than on PRN 20 may be explained, at least in part, by examining the elevation angles of the satellites during the data collection (Fig. 15.38). Lower elevation angle satellites emit signals that have greater opportunity to interact with ground objects and produce multipath. That general principle is observed here with PRN 20 in the roughly 20–30° elevation angle range whereas PRN 04 rises from slightly over 50° at the beginning of the data to nearly 80° at the end.

15.8.4 Measurement of Carrier-Phase Multipath

One technique used to identify multipath contamination on the carrier-phase is through observation of the mea-

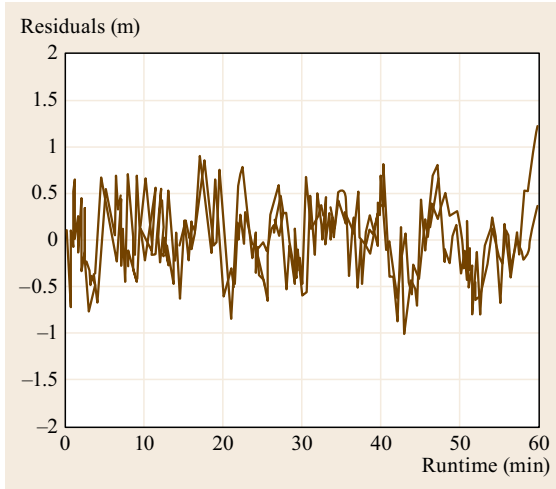


Fig. 15.37 The code-minus-carrier observables of Fig. 15.36 for all 3 days are plotted on top of each other. Multipath error during the periods from 15 to 20 min and from 40 to 50 min (run time) is clearly visible

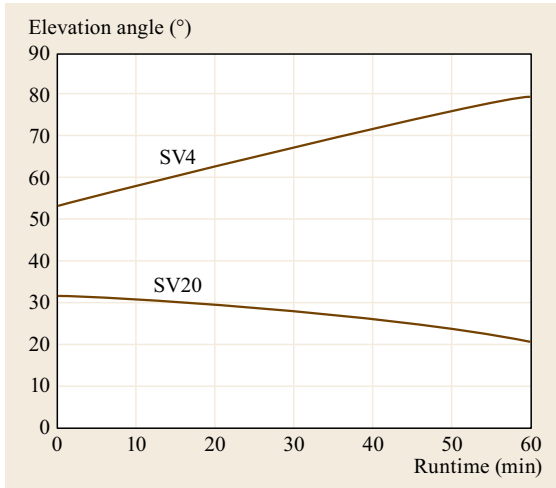


Fig. 15.38 Elevation angle for the two satellites analyzed

sured signal-to-noise (SNR) or carrier-to-noise ratio. It has been shown [15.44] that there is a high degree of correlation between oscillations in the measured SNR and carrier-phase multipath. Specifically, the larger the amplitude fluctuations in the measured SNR, the greater the impact of multipath on the carrier-phase.

As just described, the synergy between the pseudorange and carrier-phase measurements enable a pseudorange multipath observation to be constructed. However, to observe multipath on the carrier-phase measurement itself, a different combination is needed. The commonly utilized technique involves the use of two separated, but stationary, receivers tracking at least two satellites.

Furthermore, there should be negligible multipath contamination for both satellites at the reference receiver site and the test receiver site should only experience significant multipath on one of the two satellites. Under this specific scenario, it is possible to isolate the carrier-phase multipath error on the contaminated satellite measurement.

It can be shown (Chap. 20) that for closely spaced receivers the carrier-phase double-difference observation can be modeled by

$$\varphi_{km}^{pq} = \frac{1}{\lambda} R_{km}^{pq} + S_{km}^{pq} + N_{km}^{pq}, \quad (15.24)$$

where receivers k and m are tracking satellites p and q ; R is the true geometric range double difference; S is the double difference of the noise and multipath errors and N is the double-difference ambiguity. If the two receivers are in surveyed locations, the true geometric range double difference can be calculated either from broadcast ephemerides or from post-processed precise ephemerides. Obviously the precise ephemerides are preferred but if the baseline is short (i.e., less than a few kilometers), the typical errors in the broadcast ephemeris will cancel in the double differences. Assuming no cycle-slips, the double-difference ambiguity is a constant which can easily be estimated and subtracted.

After removal of the geometry and ambiguity, the result is the double difference combination of multipath and noise errors on the four constituent carrier-phase measurements. As mentioned earlier, ideally a reference receiver is established at a clean site where negligible multipath is experienced. For the second receiver, a first satellite is chosen that also has negligible multipath. The observable is then dominated by the multipath error on the second satellite [15.25]. For highly controlled laboratory experiments with carrier-phase multipath, this technique works especially well since hardware simulators can be set up to simulate zero multipath on the signals provided to the reference receiver and, similarly, multipath can be simulated on a single satellite at the second receiver. This technique was used in [15.8] to validate the carrier-phase multipath models described earlier in this chapter.

To illustrate this technique with actual field measurements, data were processed from two CORS receivers near Denver, Colorado in the United States. The data were collected between 5 am and 6 am, GPS time, on days 100 and 101 of 2014 at the ZDV1 and P041 CORS sites. The aforementioned observable was formed and the geometric range double difference was calculated and subtracted out. The results for GPS satellites PRN 4 and PRN 10 are shown in Fig. 15.39. These satellites were chosen according to the required criteria with both satellites exhibiting little multipath at

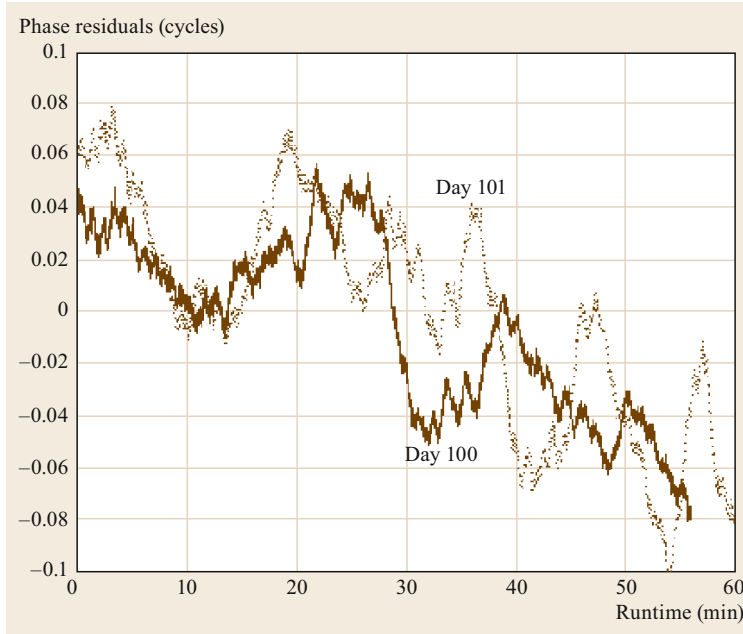


Fig. 15.39 Carrier-phase double-difference multipath residuals. The data were collected from the ZDV1 and P041 CORS sites near Denver, Colorado from 5 am to 6 am (GPS time) on days 100 and 101 of 2014. Double differences were formed with GPS satellites PRN 4 and PRN 10. The geometric range double differences were calculated, based on the surveyed locations and broadcast satellite ephemerides, and were subtracted from the double-difference observations. The result is dominated by noise and multipath. The residuals for day 100 have been shifted by 4 min to account for the sidereal day shift

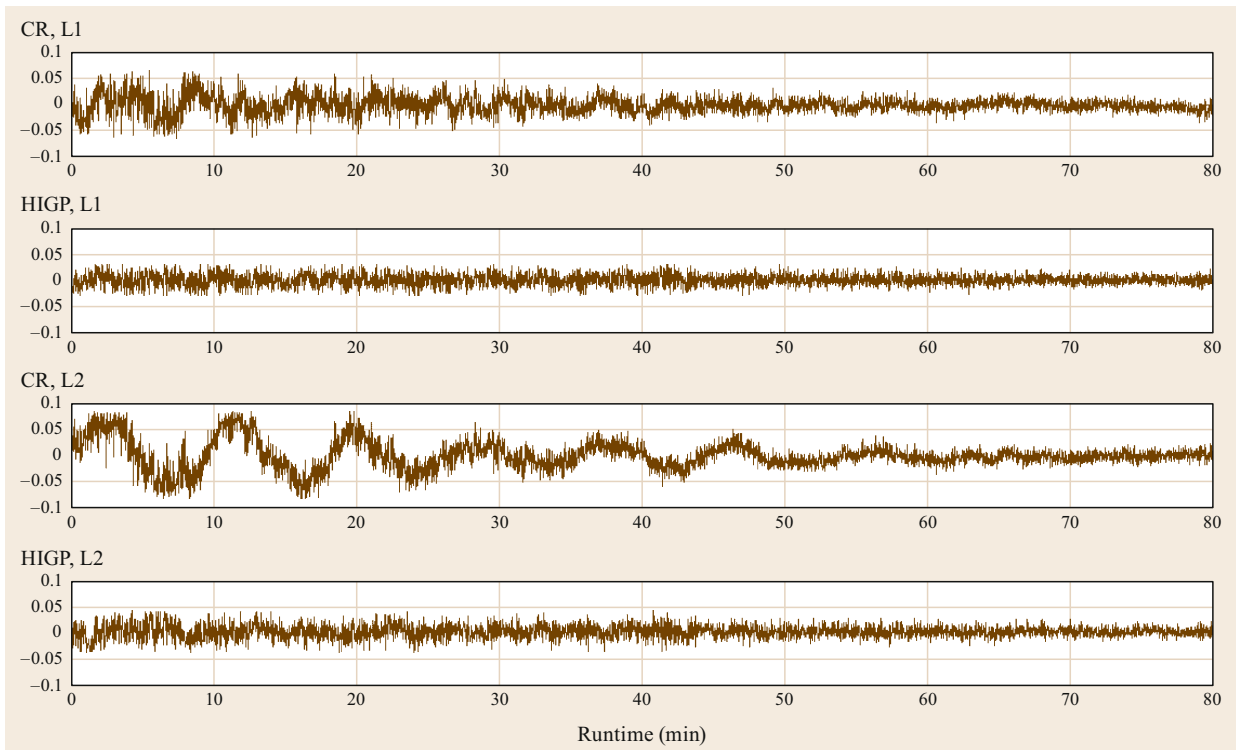


Fig. 15.40 L1 and L2 carrier-phase double-difference multipath residuals illustrating the performance of a conventional choke-ring (CR) antenna versus a novel high impedance ground plane (HIGP) antenna. The y-axes are in units of carrier wavelengths. The error oscillations due to ground reflection are clearly shown in the choke-ring results as is the efficacy of the novel antenna ground plane (courtesy of Dmitry Tatarnikov of Topcon Positioning Systems)

the P041 site and only PRN 10 exhibiting strong multipath at the ZDV1 site. The residuals from the 2 days match quite closely in the region from 5 to 15 min. This repeatability indicates that the residuals in this region are dominated by a stationary multipath source. Overall, the residuals from the two days share a common trend with a negative slope. This indicates the presence of a multipath source that is very close to the antenna and thus produces multipath with a very low frequency component. The inconsistency between the residuals of the 2 days is likely due to multipath sources that are not stationary and thus can vary from day to day (e.g., cars parked in the vicinity of one or both of the antennas).

Another example of the use of this technique is given in [15.45]. The paper describes a field data collection effort performed to determine how well a novel

high-impedance ground plane would attenuate ground reflections. The double-difference technique was used to process data over a short baseline with a choke-ring antenna as a control and with the novel antenna under study. The key results are illustrated in Fig. 15.40. The efficacy of the novel antenna ground plane is clearly illustrated. The multipath error oscillations, shown particularly well in the L2/choke-ring residuals, are due to the change of the relative delay of the ground reflection as the satellite passed through the sky. Since the antennas were mounted approximately 2 m above the ground [15.45], a fading frequency of approximately 1.5 mHz would be expected at L2 for low elevation angles (as calculated from (15.4)). This corresponds to a period of approximately 11 min which matches closely with the periods observed in the plot.

15.9 A Note About Multipath Impact on Doppler Measurements

The most accurate Doppler (LOS velocity) measurements consist of scaled, temporal differences of carrier-phase measurements. The time differences are typically 1 s or less. Thus, the error in the Doppler measurement consists of the change of the error terms over the differencing interval. Recalling the carrier-phase observable (15.19), the error terms are the satellite and clock offsets, ionospheric and tropospheric delays, multipath,

noise, dynamic tracking error, and the ambiguity. Thus the multipath component of Doppler measurement error consists of the change of carrier-phase multipath error over the time differencing interval. For cases in which the relative multipath phase-rate (fading frequency) is small (e.g., less than a hundredth of a Hertz) then the multipath component of the Doppler measurement error will be negligible.

15.10 Conclusions

In most differential GNSS (DGNSS) applications, multipath is a dominant error source. Measurement techniques can be used to assess the level of impact for a given situation. Antenna design/placement, choice of receiver architecture and measurement processing techniques all can be used to reduce multipath error. Higher rate codes for civilian use, made possible through so-called *GPS modernization* and the maturation of constellations such as Galileo will effectively eliminate the medium and long delay multipath error associated with lower rate codes.

Acknowledgments. The author would like to thank his current and former colleagues at Ohio University: Dr. Michael DiBenedetto, Dr. Sai Kalyanaraman (now with Rockwell Collins), and Mr. Joseph Kelly (now with mCube), along with Dr. Gary McGraw of Rockwell Collins, for their collaboration on various research projects that were drawn upon for material for this chapter. The author would also like to thank Dr. Dmitry Tatarnikov of Topcon Positioning Systems for the provision of the carrier-phase multipath data. Finally, the author would like to thank the US FAA, NASA, and DoD, along with Honeywell and Rockwell Collins for their support.

References

- 15.1 G. Hein, A. Teuber, H. Thierfelder, A. Wolfe: GNSS indoors: Fighting the fading – Part 2, *Inside GNSS* 3(4), 47–53 (2008)
- 15.2 K. Kaiser: Plane wave shielding. In: *Electromagnetic Compatibility Handbook*, (CRC, Boca Raton 2004) p. 21–38

- 15.3 A. Kavak, G. Xu, W. Vogel: GPS multipath fade measurements to determine L-band ground reflectivity properties, Proc. 20th NASA Propag. Exp. Meet. (NAPEX 20), Fairbanks (Jet Propulsion Laboratory, Pasadena 1996) pp. 257–263
- 15.4 P. Beckmann, A. Spizzichino: *The Scattering of Electromagnetic Waves from Rough Surfaces* (Pergamon/Macmillan, New York 1963)
- 15.5 A. Ghasemi, A. Abedi, F. Ghasemi: *Propagation Engineering in Wireless Communications* (Springer, New York 2012)
- 15.6 D. Barton: *Radar Equations for Modern Radar* (Artech House, Boston 2012)
- 15.7 R. van Nee: Spread-spectrum code and carrier synchronization errors caused by multipath and interference, IEEE Trans. Aerosp. Electron. Syst. **29**(4), 1359–1365 (1993)
- 15.8 S. Kalyanaraman, M. Braasch, J. Kelly: Code tracking architecture influence on GPS carrier phase, IEEE Trans. Aerosp. Electron. Syst. **42**(2), 548–561 (2006)
- 15.9 M. Irsigler, J. Avila-Rodriguez, G. Hein: Criteria for GNSS multipath performance assessment, Proc. ION GNSS 2005, Long Beach (ION, Virginia 2006) pp. 2166–2177
- 15.10 M. Braasch: Autocorrelation sidelobe considerations in the characterization of multipath errors, IEEE Trans. Aerosp. Electron. Syst. **33**(1), 290–295 (1997)
- 15.11 G. Brodin: GNSS code and carrier tracking in the presence of multipath, Proc. ION GPS 1996, Kansas City (ION, Virginia 1996) pp. 1389–1398
- 15.12 J. Ray: Mitigation of GPS Code and Carrier Phase Multipath Effects Using a Multi-Antenna System, Ph.D. Thesis (Univ. Calgary, Calgary 2000)
- 15.13 J. Kelly, M. Braasch, M. DiBenedetto: Characterization of the effects of high multipath phase rates in GPS, GPS Solutions **7**(1), 5–15 (2003)
- 15.14 F. Bletzacker: Reduction of multipath contamination in a geodetic GPS receiver, Proc. 1st Int. Symp. Precise Position. Glob. Position. Syst. (US Department of Commerce, National Oceanic and Atmospheric Administration, Rockville 1985) pp. 413–422
- 15.15 J. Tranquilla, J. Carr, H. Al-Rizzo: Analysis of a choke ring groundplane for multipath control in global positioning system (GPS) applications, IEEE Trans. Antennas Propag. **42**(7), 905–911 (1994)
- 15.16 W. Kunysz: A novel GPS survey antenna, Proc. ION NTM 2000, Anaheim (ION, Virginia 2000) pp. 698–705
- 15.17 W. Kunysz: High performance GPS pinwheel antenna, Proc. ION GPS 2000, Salt Lake City (ION, Virginia 2000) pp. 2506–2511
- 15.18 A. Lopez: GPS Antenna Systems, US Patent (Application) 5 534 882 (1996)
- 15.19 M. Braasch: Optimum antenna design for DGPS ground reference stations, Proc. ION GPS 1994, Salt Lake City (ION, Virginia 1994) pp. 1291–1297
- 15.20 F. van Graas, D. Diggle, M. Uijt de Haag, T. Skidmore, M. DiBenedetto, V. Wulschleger, R. Velez: Ohio Univ./FAA flight test demonstration of local area augmentation system (LAAS), Navigation **45**(2), 129–136 (1998)
- 15.21 D.B. Thornberg, D.S. Thornberg, M. DiBenedetto, M. Braasch, F. van Graas, C. Bartone: LAAS integrated multipath-limiting antenna, Navigation **50**(2), 117–130 (2003)
- 15.22 A. Lopez: GPS ground station antenna for local area augmentation system, LAAS, Proc. ION ITM 2000, Anaheim (ION, Virginia 2000) pp. 738–742
- 15.23 A. Lopez: LAAS/GBAS ground reference antenna with enhanced mitigation of ground multipath, Proc. ION NTM 2008, San Diego (ION, Virginia 2008) pp. 389–393
- 15.24 L. Garin, F. van Diggelen, J. Rousseau: Strobe and edge correlator multipath mitigation for code, Proc. ION GPS 1996, Kansas City (ION, Virginia 1996) pp. 657–664
- 15.25 L. Garin, J. Rousseau: Enhanced strobe correlator multipath rejection for code and carrier, Proc. ION GPS 1997, Kansas City (ION, Virginia 1997) pp. 559–568
- 15.26 R. Hatch, R. Keegan, T. Stansell: Leica's code and phase multipath mitigation techniques, Proc. ION NTM 1997, Santa Monica (ION, Virginia 1997) pp. 217–225
- 15.27 L. Weill: GPS multipath mitigation by means of correlator reference waveform design, Proc. ION NTM 1997, Santa Monica (ION, Virginia 1997) pp. 197–206
- 15.28 L. Weill: Application of superresolution concepts to the GPS multipath mitigation problem, Proc. ION NTM 1998, Long Beach (ION, Virginia 1998) pp. 673–682
- 15.29 G. McGraw, M. Braasch: GNSS multipath mitigation using gated and high resolution correlator concepts, Proc. ION NTM 1999, Santa Diego (ION, Virginia 1999) pp. 333–342
- 15.30 M. Irsigler, B. Eissfeller: Comparison of multipath mitigation techniques with consideration of future signal structures, Proc. ION GPS 2003, Portland (ION, Virginia 2003) pp. 2584–2592
- 15.31 J. Jones, P. Fenton, B. Smith: *Theory and Performance of the Pulse Aperture Correlator* (NovAtel, Calgary 2004), <http://www.novatel.com/assets/Documents/Papers/PAC.pdf>
- 15.32 T. Pany, M. Irsigler, B. Eissfeller: S-curve shaping: A new method for optimum discriminator based code multipath mitigation, Proc. ION GNSS 2005, Long Beach (ION, Virginia 2005) pp. 2139–2154
- 15.33 M. Paonni, J. Avila-Rodriguez, T. Pany, G. Hein, B. Eissfeller: Looking for an optimum S-curve shaping of the different MBOC implementations, Navigation **55**(4), 255–266 (2008)
- 15.34 Y. Bock: Continuous monitoring of crustal deformation, GPS World **2**(6), 40–47 (2008)
- 15.35 K. Choi, A. Bilich, K. Larson, P. Axelrad: Modified sidereal filtering: Implications for high-rate GPS positioning, Geophys. Res. Lett. **31**(L22608), 1–4 (2004)
- 15.36 Z. Zhu, S. Gunawardena, M. Uijt de Haag, F. van Graas: Satellite anomaly and interference detection using the GPS anomalous event monitor, Proc. ION AM 2007, Cambridge (ION, Virginia 2007) pp. 389–396

- 15.37 Z. Zhu, S. Gunawardena, M. Uijt de Haag, F. van Graas: Advanced GPS performance monitor, Proc. ION GNSS 2007, Fort Worth (ION, Virginia 2007) pp. 415–423
- 15.38 M. Irsigler: Multipath Propagation, Mitigation and Monitoring in the Light of Galileo and the Modernized GPS, Ph.D. Thesis (Universität der Bundeswehr, Munich 2008)
- 15.39 Z. Zhu, S. Gunawardena, M. Uijt de Haag, F. van Graas, M. Braasch: GNSS watch dog: A GPS anomalous event monitor, Inside GNSS **3**(7), 18–28 (2008)
- 15.40 Final User Field Test Report for the NAVSTAR Global Positioning System Phase I, Major Field Test Objective No. 17: Environmental Effects, Multipath Rejection, Rep. GPS-GD-025-C-US-7008 (General Dynamics Electronics Division, San Diego 1979)
- 15.41 A. Evans: Comparison of GPS pseudorange and biased doppler range measurements to demonstrate signal multipath effects, Proc. Int. Telemetering Conf., Las Vegas (1986) pp. 795–801
- 15.42 M. Braasch: Isolation of GPS multipath and receiver tracking errors, Navigation **41**(4), 415–434 (1994)
- 15.43 R. Langley: GPS receivers and the observables. In: *GPS for Geodesy*, Vol. 2, ed. by P. Teunissen, A. Kleusberg (Springer, Berlin 1998) pp. 167–171
- 15.44 P. Axelrad, C. Comp, P. MacDoran: SNR-based multipath error correction for GPS differential phase, IEEE Trans. Aerosp. Electron. Syst. **32**(2), 650–660 (1996)
- 15.45 D. Tatarnikov, A. Astakhov: Approaching millimeter accuracy of GNSS positioning in real time with large impedance ground plane antennas, Proc. ION NTM 2014, Santa Diego (ION, Virginia 2014) pp. 844–848

Interference

16. Interference

Todd Humphreys

Global navigation satellite system (GNSS) signals are so weak near the Earth's surface that they can be easily squelched by natural or man-made interference. Moreover, the most popular GNSS signals – those offered with unrestricted access – are unencrypted and unauthenticated, which means they can be counterfeited, or spoofed. Strict international laws protect the radio frequency bands allocated to GNSS, but mother nature does not respect these laws, and man-made interference – whether accidental or intentional – is a growing concern.

This chapter examines sources of GNSS signal interference and the interference effects on GNSS signal tracking. It offers a systematic treatment of natural, unintentional, and intentional interference, with emphasis on intentional jamming and spoofing. Theoretical performance bounds are developed for the simplest cases of narrowband and wideband interferences. The chapter finishes with a review of the state of the art in antenna-oriented and signal-processing-oriented interference detection and mitigation techniques.

16.1	Analysis Technique for Statistically Independent Interference	471
16.1.1	Received Signal Model	471
16.1.2	Thermal-Noise-Equivalent Approximation	471
16.1.3	Limits of Applicability	473
16.1.4	Overview of Interference Effects on Carrier Phase Tracking	474
16.2	Canonical Interference Models	476

16.2.1	Wideband Interference	476
16.2.2	Narrowband Interference	476
16.2.3	Matched-Spectrum Interference	478
16.3	Quantization Effects	479
16.3.1	One-Bit Quantization	479
16.3.2	Multibit Quantization	479
16.4	Specific Interference Waveforms and Sources	481
16.4.1	Solar Radio Bursts	481
16.4.2	Scintillation	482
16.4.3	Unintentional Interference	484
16.4.4	Intentional Interference	485
16.5	Spoofing	485
16.5.1	Generalized Model for Security-Enhanced GNSS Signals	486
16.5.2	Attacks Against Security-Enhanced GNSS Signals	486
16.6	Interference Detection	491
16.6.1	C/N_0 Monitoring	491
16.6.2	Received Power Monitoring	491
16.6.3	Augmented Received Power Monitoring	493
16.6.4	Spectral Analysis	494
16.6.5	Cryptographic Spoofing Detection	495
16.6.6	Antenna-Based Techniques	497
16.6.7	Innovations-Based Techniques	497
16.7	Interference Mitigation	498
16.7.1	Spectrally or Temporally Sparse Interference	498
16.7.2	Spectrally and Temporally Dense Interference	499
16.7.3	Antenna-Based Techniques	500
	References	501

All GNSS waveforms are spread-spectrum signals, which are uniquely resilient to interference. Indeed, robustness in the face of jamming was one of the primary features, along with low probability of intercept and good multiple access properties, which motivated the original development of spread-spectrum techniques for military systems. Nonetheless, GNSS signals are extremely vulnerable to jamming because, near the sur-

face of Earth, they have no more flux density than light received from a 50 W bulb at a distance of 2000 km. To blandly remark that GNSS signals are weak is to understate their fragility: They are so weak that most modern electronics jam GNSS receivers at close range, requiring special precautions be taken to isolate receivers embedded in computers, mobile phones, vehicles, and other modern GNSS-dependent systems.

Table 16.1 ITU space-to-Earth radio navigation satellite service (RNSS) frequency allocations (after [16.2, 3]). ARNS refers to the Aeronautical Radionavigation Service. Bands that are designated as both RNSS and ARNS enjoy, in principle, no greater International Telecommunication Union (ITU) protection from harmful interference than RNSS bands, but in practice they are granted more conservative safety margins (see, e.g., ITU-R M.1903) and they are likely to be monitored more assiduously by ITU member nations

Frequency interval (MHz)	Bandwidth (MHz)	GNSS bands	Notes
1164–1215	51	L5/E5a/E5b/L3/B2	ARNS band; pulsed DME/TACAN interference present [16.1]
1215–1240	25	L2	Legacy GPS L2 band
1240–1260	20	L2	Legacy GLONASS L2 band
1260–1300	40	E6/B3/LEX	
1559–1610	51	L1/E1/B1	ARNS band; legacy GPS and GLONASS L1 band
5010–5030	20	C1	

Unintentional and intentional GNSS interferences are distinguished from each other more by motive than by effect. Both can be narrowband or wideband (relative to the bandwidth of the desired GNSS signal), structured or random. The user of a GNSS receiver suffering from interference may care little about the jammer's intent: What is important is a clean spectrum. Indeed, the recent emergence of so-called personal privacy devices (PPDs) – low-cost GNSS jammers used to ward off GNSS tracking – blurs the lines between unintentional and intentional interference: The privacy device user only intends to jam GNSS receivers in an imaginary bubble around himself; he may never intend to disrupt the GNSS-dependent timing system at the bank down the street.

Interference that mimics GNSS signal structure and content is a special threat to GNSS receivers. Instead of simply degrading the accuracy of the position, velocity, and time (PVT) solution, transmission of such structured interference, referred to as spoofing, can fool a receiver into producing a precise but erroneous solution. Worse yet, the induced solution can be entirely dictated by the spoofer operator, who may have malevolent intentions. All GNSS signals are spoofable to one degree or another – at the very least, they can all be recorded and replayed into a target receiver, as is routinely done for receiver testing. But the most popular GNSS signals, the so-called open signals, are especially vulnerable because they are (so far) almost entirely predictable, lacking encryption or authentication of any form. For radionavigation as for communication, predictability is the enemy of security.

From the origins of GNSS, national and international policy has afforded special protection to the GNSS radio bands, and now that GNSS receivers have become pervasively embedded in the infrastructure that supports the global economy, such protection is of spe-

cial importance. The International Telecommunication Union (ITU) forbids any interference *which endangers the functioning of a radionavigation service* [16.2] in the GNSS bands, which are designated as radionavigation satellite service (RNSS) bands by the ITU. Table 16.1 summarizes the ITU's current frequency allocations for GNSS signals.

In some regions, the penalty for emitting unauthorized signals in the GNSS bands is severe: In response to a rising number of so-called PPDs, the United States Federal Communications Commission (FCC) levies costly fines on intentional violators [16.4], and the penalty for intentional transmission in Australia can include a 2 yr prison term [16.5]. But despite government protections of the GNSS bands, they remain cluttered with interference, and there is every indication that such interference will worsen in the decades to come as more GNSS constellations begin broadcasting [16.6], as people respond to pervasive GNSS tracking by employing PPDs [16.7], and as communications signals ineluctably encroach on the enormously valuable GNSS spectral bands [16.8].

This chapter examines the effects of interference on GNSS receivers. The chapter begins with a presentation of the general analysis technique that will be used to evaluate the effect of interference that is statistically independent of the GNSS signals. The technique will then be applied to study the effects of canonical narrowband, wideband, and multiaccess interference. Following this, other specific interference waveforms such as pulsed interference will be discussed. Thereafter, GNSS spoofing, a particular type of interference that cannot be considered statistically independent of the GNSS signals, will be given a focused treatment. The chapter finishes with an examination of interference detection and mitigation strategies. Note that GNSS multipath, while a genuine type of interference, is treated separately in Chap. 15.

16.1 Analysis Technique for Statistically Independent Interference

Beyond the statement that GNSS interference always degrades PVT accuracy, one can say little in general about interference effects on late-stage signal processing products because these effects are highly receiver-dependent: A vector-tracking low-tracking-bandwidth receiver will, for example, produce a much more robust PVT solution than a scalar-tracking wide-bandwidth receiver. At earlier processing stages, however, interference effects are substantially common across receiver types and thus a general treatment becomes possible. Accordingly, this section presents an analysis of interference effects on the primitive correlation-and-accumulation products that form the basis of signal tracking in all GNSS receivers.

16.1.1 Received Signal Model

Consider the following generic representation of a received GNSS signal exiting a receiver's radio frequency (RF) front-end downconversion chain. For notational compactness, the signal is expressed by its complex baseband representation as

$$r_S(t) = \sqrt{P_S} D(t - \tau(t)) C(t - \tau(t)) \exp(j\theta(t)), \quad (16.1)$$

where P_S is the received signal power in watts, $D(t)$ is the binary navigation data modulation, $C(t)$ is the binary spreading (ranging) code, $\tau(t)$ is the code phase, and $\exp(j\theta(t))$ is the carrier with phase $\theta(t)$. The code phase $\tau(t)$ varies slowly and, for purposes of interference modeling and analysis, can be modeled as constant; thus, it will be denoted τ hereafter.

Let $r_I(t)$ represent a complex-valued interference signal, and let $n(t) = n_I(t) + jn_Q(t)$ be a zero-mean complex-valued Gaussian process that models thermal noise. Then, the full received signal-plus-interference-and-noise is given by

$$r(t) = r_S(t) + r_I(t) + n(t).$$

The received components $r_S(t)$, $r_I(t)$, and $n(t)$ are assumed to be limited by a bandpass filter in the RF front end having a noise-equivalent bandwidth of W_{FE} Hz. The quadrature processes $n_I(t)$ and $n_Q(t)$ are modeled as spectrally flat on the range, $|f| < W_{FE}/2$ with two-sided density $N_0/2$, where N_0 has units of W/Hz. Consequently, on this range the full complex thermal noise process $n(t)$ has a two-sided density of N_0 . The data $D(t)$ and spreading code $C(t)$ are assumed to be nor-

malized to unity power so that

$$P_S = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} |r_S(t)|^2 dt.$$

If $r_S(t)$, $r_I(t)$, and $n(t)$ are statistically independent, then the total received power in the bandwidth W_{FE} , denoted by P_T , is

$$P_T = P_S + P_I + P_n, \quad (16.2)$$

where P_I is the total power in $r_I(t)$, and $P_n = W_{FE}N_0$. The carrier power to thermal-noise density ratio is $C/N_0 = P_S/N_0$, and the signal-to-thermal-noise ratio is $\text{SNR}_{FE} = P_S/P_n$. Similarly, the signal-to-interference-and-thermal-noise ratio is $\text{SINR}_{FE} = P_S/(P_n + P_I)$. Figure 16.1 offers an example illustration of the relationship between the power spectra of $r_S(t)$, $r_I(t)$, and $n(t)$.

16.1.2 Thermal-Noise-Equivalent Approximation

A key insight greatly simplifies GNSS interference analysis: The effect of interference on almost all GNSS receiver functions can be accurately modeled as if it were caused by spectrally flat thermal noise of a certain density. This subsection explains when this thermal-noise-equivalent approximation is valid and notes its limitations.

GNSS signal processing is founded on correlation of the received signal $r(t)$ with a local replica

$$l(t) = C_I(t - \hat{\tau}) \exp(j\hat{\theta}(t)),$$

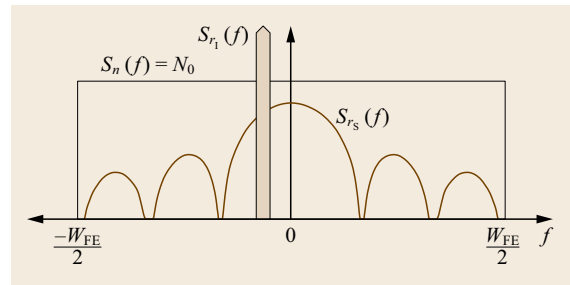


Fig. 16.1 Stylized depiction of the power spectra $S_{r_S}(f)$, $S_{r_I}(f)$, and $S_n(f)$ that correspond, respectively, to the received components $r_S(t)$, $r_I(t)$, and $n(t)$. The spectra are assumed to be significant only within the interval $|f| \leq W_{FE}/2$, where W_{FE} is the bandwidth of the RF front end's narrowest bandpass filter. The total power in $S_{r_S}(f)$, $S_{r_I}(f)$, and $S_n(f)$ within this interval is, respectively, P_S , P_I , and P_n .

where, ignoring the effects of band-limiting, $C_I(t)$ is often taken to be equal to $C(t)$, though it may differ from $C(t)$ when modeling early-minus-late correlation or when a specialized code replica is generated to reduce multipath. Suppose that a GNSS receiver is tracking the carrier phase of $r_S(t)$ so that $\hat{\theta}(t) \approx \theta(t)$. Then, the complex correlator output

$$Y(t) \equiv r^*(t)l(t) = S(t) + I(t) + N(t) \quad (16.3)$$

is composed of the desired component

$$S(t) \approx \sqrt{P_S} D(t - \tau) C(t - \tau) C_I(t - \hat{\tau}),$$

an interference component

$$I(t) = r_I^*(t) C_I(t - \hat{\tau}) \exp(j\hat{\theta}(t)),$$

and a random noise component $N(t) = n^*(t)l(t)$.

If the components $r_I^*(t)$, $C_I(t - \hat{\tau})$, and $\exp(j\hat{\theta}(t))$ are wide-sense stationary and mutually statistically independent, as is a reasonable approximation for non-spoofing interference, then the autocorrelation function of $I(t)$ can be expressed as

$$\begin{aligned} R_I(\tilde{\tau}) &\equiv E[I^*(t)I(t - \tilde{\tau})] \\ &= E[r_I^*(t)r_I^*(t - \tilde{\tau})] \\ &\quad \times E[C_I(t - \hat{\tau})C_I(t - \tilde{\tau} - \hat{\tau})] \\ &\quad \times E(\exp(j\hat{\theta}(t)) \exp(j\hat{\theta}(t - \tilde{\tau}))) . \end{aligned} \quad (16.4)$$

In other words, $R_I(\tilde{\tau})$ is the product of the autocorrelation functions corresponding to each of the three components of $I(t)$. Consequently, the power spectral density of $I(t)$, $S_I(f) = \mathcal{F}[R_I(\tilde{\tau})]$, where \mathcal{F} denotes the Fourier transform, can be found by convolving the power spectra of the three components. Let $S_{C_I}(f)$, $S_{r_I}(f)$, and $\delta(f + \hat{f}_D)$ be the respective power spectra of $C_I(t)$, $r_I(t)$, and $\exp(j\hat{\theta}(t))$, where

$$\hat{f}_D = -\frac{1}{2\pi} \frac{d\hat{\theta}}{dt}$$

is the receiver's estimate of the desired signal's apparent Doppler frequency, in Hz, and $\delta(f)$ is the Dirac delta function. It follows that

$$\begin{aligned} S_I(f) &= S_{C_I}(f) * S_{r_I}(f) * \delta(f + \hat{f}_D) \\ &= S_{C_I}(f) * S_{r_I}(f + \hat{f}_D) , \end{aligned}$$

where $*$ denotes convolution.

The values of $S_I(f)$ within a narrow neighborhood about $f = 0$ are a useful starting point for predicting GNSS interference effects. To understand why, consider the block diagram in Fig. 16.2, which illustrates correlation of the received signal $r(t)$ with the local signal replica $l(t)$ followed by an accumulate-and-dump operation that produces the discrete complex accumulation products $Y_k = I_k + jQ_k$, $k = 1, 2, \dots$. The accumulate-and-dump operation acts as a low-pass filter having a squared frequency response

$$|H_a(f)|^2 = \text{sinc}^2(fT_a) ,$$

where $\text{sinc}(x) \equiv \sin(\pi x)/\pi x$ and T_a is the accumulation interval in seconds. The interference power that passes through the accumulate-and-dump filter into the complex accumulation products – and thereafter into the code and carrier tracking loops – is given by

$$P_{\text{al}} = \int_{-\infty}^{\infty} |H_a(f)|^2 S_I(f) df .$$

Let the noise-equivalent bandwidth of the accumulate-and-dump filter be defined as

$$W_a \equiv \int_{-\infty}^{\infty} \text{sinc}^2(fT_a) df = \frac{1}{T_a}$$

and let $I_0 \equiv S_I(0)$. Then, so long as $S_I(f)$ is nearly constant (flat) over a few multiples of W_a , P_{al} can be approximated as

$$P_{\text{al}} \approx \tilde{P}_{\text{al}} \equiv I_0 W_a .$$

For typical values of T_a , and for typical spreading code replicas $C_I(t)$, the quasi-constant condition on $S_I(f)$ is easily satisfied. To understand why, consider Fig. 16.3 in connection with the following argument. Assume that $S_{C_I}(f)$ and $S_I(f)$ are smooth (no spectral lines) with respective frequency derivatives $S'_{C_I}(f)$ and $S'_I(f)$. The error in the approximating P_{al} by \tilde{P}_{al} can be expressed in dB as

$$\Delta P_{\text{al}} \equiv 10 \log_{10} \left(\left| \frac{\tilde{P}_{\text{al}}}{P_{\text{al}}} \right| \right) ,$$

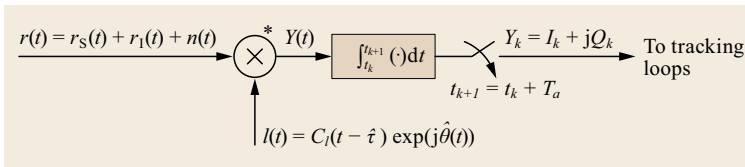


Fig. 16.2 Block diagram of the standard correlation and accumulation process in a GNSS receiver. The complex product of the incoming signal $r(t)$ and the local replica $l(t)$ is accumulated over T_a seconds to produce the discrete complex-valued accumulation product Y_k

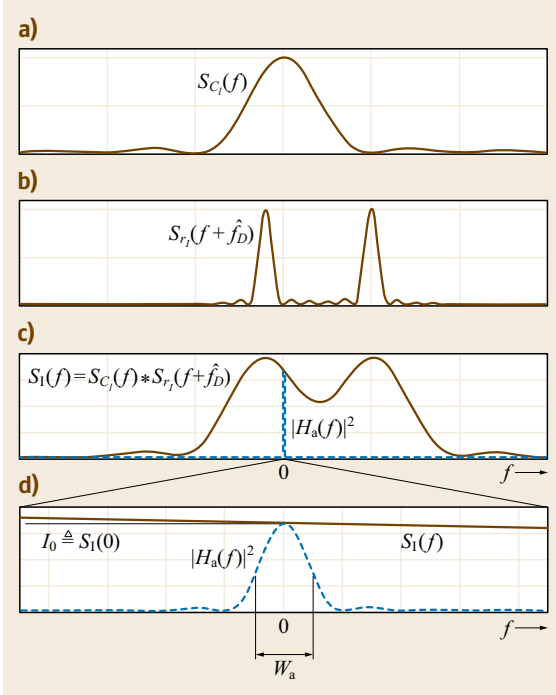


Fig. 16.3a–d Example power spectra and filtering involved in interference analysis: (a) $S_{C_l}(f)$, the spectrum of the GNSS replica code; (b) $S_{r_l}(f + \hat{f}_D)$, the spectrum of the received interference convolved with $\delta(f + \hat{f}_D)$; (c) $S_I(f)$, the spectrum of $I(t)$, together with $|H_a(f)|^2$, the squared frequency response of the accumulate-and-dump filter; (d) zoomed view of $S_I(f)$ and $|H_a(f)|^2$ near $f = 0$ showing that, despite the interference being fairly narrowband, $S_I(f)$ is approximately flat over the noise-equivalent bandwidth W_a

which for practical $C_l(t)$ satisfies

$$\Delta P_{al} < 10 \log_{10} \left(1 + \frac{|S'_{C_l}(0)|}{S_{C_l}(0)} W_a \right).$$

But, due to the properties of convolution,

$$\frac{|S'_{C_l}(0)|}{S_{C_l}(0)} \leq \max_f \frac{|S'_{C_l}(f)|}{S_{C_l}(f)}.$$

And note that when performing the maximization, one need only consider f values within

$$\mathcal{U}_\epsilon = \{f | \epsilon < S_{C_l}(f)\}$$

for some $\epsilon > 0$ because for $f \notin \mathcal{U}_\epsilon$ the possible contribution of $|S'_{C_l}(f)|/S_{C_l}(f)$ to P_{al} is small, making large values of $|S'_{C_l}(f)|/S_{C_l}(f)$ immaterial. Putting these

pieces together, ΔP_{al} can be upper bounded as

$$\Delta P_{al} < \max_{f \in \mathcal{U}_\epsilon} \left[10 \log_{10} \left(1 + \frac{|S'_{C_l}(f)|}{S_{C_l}(f)} W_a \right) \right].$$

Consider an example designed for large ΔP_{al} . Let $C_l(t)$ be matched to the relatively narrowband GPS L1 C/A code (ignoring spectral lines), for which

$$S_{C_l}(f) = T_C \text{sinc}^2(fT_C)$$

$$S'_{C_l}(f) = \frac{2T_C}{f} [\text{sinc}(2fT_C) - \text{sinc}^2(fT_C)],$$

where $T_C \approx 1 \mu\text{s}$ is the spreading code chip interval. Choosing $\epsilon = S_{C_l}(0)/100$, it can be shown that $|S'_{C_l}(f)/S_{C_l}(f)|$ achieves a maximum of approximately $25T_C$ so that, even assuming $W_a = 1 \text{ kHz}$ – the widest typical accumulate-and-dump bandwidth for the GPS L1 C/A signal – the ratio ΔP_{al} , and thus the error in approximating P_{al} by \bar{P}_{al} , remains less than 0.105 dB, which can be considered insignificant for most applications.

The thermal-noise-equivalent approximation to interference effects can be summarized as follows. At the input to the low-pass accumulate-and-dump filter that produces the complex accumulations $Y_k = I_k + jQ_k$, the carrier-power to thermal-noise density ratio is $C/N_0 = P_S/N_0$; at the output of the filter, the signal-to-thermal noise ratio is $\text{SNR} = P_S/N_0 W_a$. When, in addition to thermal noise, interference is present, then at the filter input the carrier-power to interference-and-thermal-noise ratio (CINR) can be approximated as

$$\text{CINR} = \frac{C}{N_{0,\text{eff}}} = \frac{P_S}{N_0 + I_0},$$

where $N_{0,\text{eff}} \equiv N_0 + I_0$ is the effective thermal noise density, which accounts for both thermal noise and interference. At the filter output, the signal-to-interference-and-thermal-noise ratio can be approximated as

$$\text{SINR} = \frac{P_S}{N_{0,\text{eff}} W_a}.$$

Thus, apart from the limitations described below, analysis of GNSS receiver behavior in the presence of interference can proceed just as analysis of receiver behavior in the presence of thermal noise, which is well understood [16.9–11], by substituting CINR (or $C/N_{0,\text{eff}}$) for C/N_0 , and SINR for SNR.

16.1.3 Limits of Applicability

Approximating interference that is statistically independent of the code and carrier replicas as if it were

thermal noise with spectral density I_0 at the input of the accumulate-and-dump filter yields excellent agreement with the full theoretical error statistics for acquisition, carrier tracking, and data demodulation [16.12]. The approximation is also accurate for predicting the statistics of any coherent correlation with code replica $C_l(t)$. For example, it accurately predicts the statistics of the coherent early-minus-late code phase error so long as data bits are estimated correctly, and $C_l(t)$ is taken to be the difference between early and late code replicas [16.12]. But the thermal-noise-equivalent approximation is known to produce biased code phase error statistics for noncoherent code phase discriminators [16.13, 14]. In this case, narrowband interference maximizes code tracking error not when the interference is centered at $f = 0$ Hz (i.e., when aligned with the desired signal's carrier frequency), as one would expect, but rather when it is centered at $f \approx 1/T_C$ Hz. However, if one properly accounts for squaring loss, then even the noncoherent phase error statistics can be reduced to an accurate thermal-noise-equivalent representation [16.12]. In short, the thermal-noise-equivalent approximation has wide applicability for analysis of interference effects.

It is worth noting that if the received interference $r_1(t)$ is not statistically independent of $C_l(t - \hat{\tau})$ and $\exp(j\hat{\theta}(t))$, then factorization of $R_l(\tilde{\tau})$ as in (16.4) is not possible and the thermal-noise-equivalent approximation is not valid. This case arises, for example, when the interference is structurally similar to the desired signal $r_s(t)$ and is approximately code-phase aligned with $r_s(t)$ – in other words, when the interference is a spoofing signal. For this reason, spoofing-type interference will be treated separately later in this chapter; meanwhile, all $r_1(t)$ will be assumed to be independent of $C_l(t - \hat{\tau})$ and $\exp(j\hat{\theta}(t))$. Furthermore, all code and carrier-phase measurements will be assumed to be produced by coherent phase discriminators. Under these conditions, the thermal-noise-equivalent approximation whereby CINR is substituted for C/N_0 can be expected to accurately predict receiver effects.

16.1.4 Overview of Interference Effects on Carrier Phase Tracking

Assuming the thermal-noise-equivalent approximation to be valid, this subsection gives an overview of interference effects on carrier-phase tracking. Attention is focused on phase tracking because the phase-tracking loop, or phase lock loop (PLL), is the weakest link in the signal tracking chain. Typically, if the PLL can maintain lock, then a frequency-tracking loop and a code-phase-tracking loop can as well.

Phase Error Variance

Consider a standard (nonsquaring) PLL with true phase input $\theta(t)$ and phase estimate $\hat{\theta}(t)$. When the phase error $\varphi(t) = \theta(t) - \hat{\theta}(t)$ is small enough that the PLL's phase detector can be regarded as linear, then, for zero-mean white driving noise, the PLL's phase error variance $\sigma_\varphi^2 = E[\varphi^2(t)]$ (in rad^2) is accurately approximated by [16.15]

$$\sigma_\varphi^2 = \frac{B_n N_0}{C} \equiv \frac{1}{\rho_L}, \quad (16.5)$$

where B_n is the PLL's single-sided noise bandwidth and ρ_L is the loop SNR. GNSS carrier-phase tracking of data-modulated signals requires a squaring (e.g., Costas) PLL, which is insensitive to the half-cycle phase changes induced by the data modulation. In a squaring PLL, the actual phase error tracked is 2φ , with the corresponding variance denoted by $\sigma_{2\varphi}^2$. Furthermore, ρ_L is reduced by a squaring loss factor approximately equal to [16.16]

$$S_L = \left(1 + \frac{N_0}{2T_a C}\right)^{-1},$$

where $1/T_a$ is the predetection bandwidth. Thus, for the squaring loop,

$$\sigma_\varphi^2 = \frac{\sigma_{2\varphi}^2}{4} = \frac{1}{\rho_L S_L}$$

is a useful approximation for σ_φ^2 in the linear regime. For analysis of the squaring loop, an equivalent loop SNR is defined as [16.17, p. 206]

$$\rho_{\text{eq}} \equiv \frac{\rho_L S_L}{4}, \quad (16.6)$$

which leads to $\rho_{\text{eq}} \approx 1/\sigma_{2\varphi}^2$ for small φ .

At large values of φ , the assumption of PLL linearity breaks down and analysis becomes more difficult. An exact expression for σ_φ^2 for a first-order nonsquaring PLL driven by white Gaussian noise is found in [16.18, Chap. 4]. Precise phase error statistics for all but this standard first-order loop are typically obtained via simulation. Fortunately, one can show that the exact phase error variance for the standard first-order loop is a reasonable proxy for that of higher-order loops. Thus, one can identify the region of approximate linear PLL operation by noting that, for the standard first-order loop, the linear model in (16.5) is reasonably accurate (within 20%) for $\rho_L > 4$, or $\sigma_\varphi < 28.6^\circ$ [16.18, Chap. 4]. Likewise, a squaring loop behaves approximately linearly for $\rho_{\text{eq}} > 4$, or $\sigma_\varphi < 14.3^\circ$.

Cycle Slipping

A PLL's phase detector is periodic, meaning that it cannot distinguish between the phase errors φ and $\varphi + 2n\pi$ (nonsquaring loop) or φ and $\varphi + n\pi$ (squaring loop), where n is an integer. As a result, an infinite set of stable attractors exists for the nonlinear difference equations that describe the PLL error dynamics. At low loop SNR, the phase error can slip from one stable attractor to another, leading to infinite σ_φ^2 in the steady state. This is the familiar cycle slip phenomenon associated with PLLs [16.19, 20], [16.15, Chap. 6].

The mean time to first cycle slip T_s is defined as the average time required for the loop phase error to reach $\pm 2\pi$ ($\pm\pi$ for the squaring loop) for the first time, starting from an initial condition of zero phase error. For first-order loops, and in other cases where cycle slips occur as isolated events, T_s is the same as the mean time between cycle slips; if cycle slips occur in bursts – as may happen for $\rho_L, \rho_{eq} < 5$ in second- or higher-order loops – then T_s and the mean time between cycle slips are not related simply [16.20].

As with the calculation of σ_φ^2 , an analytical solution for T_s has only been possible for the simple case of a first-order unstressed (zero static phase error) PLL driven by white Gaussian noise, in which case [16.18, p. 101]

$$T_s = \frac{\pi^2 \rho_L I_0^2(\rho_L)}{2B_n} \quad (16.7)$$

is the time to first slip/mean time between slips for a nonsquaring loop, $I_0(\cdot)$ being a modified Bessel function of the first kind. An approximate T_s for first-order squaring loops is obtained by substituting ρ_{eq} for ρ_L . Unstressed second- and higher-order loops have lower T_s than unstressed first-order loops, and stressed loops are more prone to cycle slipping than unstressed loops; nonetheless, (16.7) remains a useful upper bound. For GNSS applications, a second- or third-order loop is required to accurately track carrier-phase in the presence of Doppler-induced quadratic phase growth. In fact, even the second-order loop experiences significant loop stress ($\approx 1^\circ$ static phase error) during the largest GNSS line-of-sight accelerations. Only the third-order loop maintains near-zero static phase error for all GNSS geometries.

Frequency Unlock

The general term *phase unlock* refers to single or successive cycle slips. At very low loop SNR, a PLL may never recover phase lock after a long succession of cycle slips. This phenomenon, called *drop lock* in the

PLL literature, is related to the PLL's frequency pull-in range. For reasons that will become clear, the term *frequency unlock* is a more precise descriptor than drop lock for the phenomenon as it relates to the discrete-time PLLs used in modern GNSS receivers.

A PLL's frequency pull-in range is the maximum frequency step input that a PLL is able to *pull in* and eventually achieve phase lock. For example, a continuous-time first-order nonsquaring PLL has a pull-in range equal to the loop gain K [16.19]. For higher-order PLLs, the frequency pull-in range can be thought of as the maximum tolerable mismatch $\Delta\omega = |\omega_c - v|$ between the carrier frequency ω_c and the PLL's internal estimate of carrier frequency v , assuming that higher-order loop filter states (e.g., the estimate of carrier frequency rate) are relaxed, where applicable.

Continuous-time PLLs whose loop filters contain one or more perfect integrators have an infinite frequency pull-in range [16.15, Chap. 8]. On the other hand, the frequency pull-in range of second- and higher-order discrete-time PLLs is limited by the loop update (accumulation) interval T_a . When the frequency mismatch $\Delta\omega$ exceeds a certain threshold $\Delta\omega_m$, then v is attracted toward a stable equilibrium value that satisfies $T_a\Delta\omega = n\pi$ (nonsquaring loop) or $T_a\Delta\omega = n\pi/2$ (squaring loop), $n = 1, 2, 3, \dots$. Intuitively, these equilibrium values exist because the loop cannot detect a phase error change of $2n\pi$ (nonsquaring loop) or $n\pi$ (squaring loop) between loop updates. The value of $\Delta\omega_m$ is a function of the particular loop configuration. It can be surprisingly small for PLLs common in GNSS receivers: for a third-order Costas loop with $T_a = 10$ ms and $B_n = 10$ Hz, $\Delta\omega_m = 81$ rad/s ≈ 13 Hz. At very low loop SNR, cycle slips can occur in bursts as noise and phase dynamics force v momentarily away from ω_c [16.20]. If, due to such forcing, $\Delta\omega$ exceeds $\Delta\omega_m$, then there is a high probability that v will become trapped at one of the incorrect stable equilibrium values. Thus, the PLL experiences frequency unlock.

Frequency unlock and momentary phase unlock have rather different practical consequences. Unlike momentary phase unlock (i.e., cycle slipping), frequency unlock often leads to complete loss of the GNSS signal link – a result of signal attenuation due to frequency detuning. If v settles on an equilibrium value such that $n \geq 2$ (nonsquaring loop) or $n \geq 4$ (squaring loop), then the baseband signal power drops by more than 13 dB, making it likely that the PLL will experience further frequency detuning and eventually lose the signal entirely. Worse yet, re-acquisition may not be possible at low SNR.

16.2 Canonical Interference Models

16.2.1 Wideband Interference

The simplest variants of $r_I(t)$ are the extreme cases of wideband and narrowband interferences. Consider first wideband interference. Suppose that $r_I(t)$ is spectrally flat with power density $S_I(f) = P_I/W_{FE}$ over a two-sided front-end bandwidth $W_{FE} \gg 1/T_C$, where T_C is the chip interval of $C(t)$ (e.g., $1/T_C = 1.023$ MHz for the GPS L1 C/A code). In this case, $S_I(f) = S_{C_I}(f) * S_{r_I}(f + \hat{f}_D) \approx S_{r_I}(f) = P_I/W_{FE}$, which implies that $I_0 \equiv S_I(0) = P_I/W_{FE}$. Hence, post-correlation error analysis can proceed by approximating the carrier-to-noise ratio as

$$\text{CINR} = \frac{C}{N_{0,\text{eff}}} = \frac{P_S}{N_0 + P_I/W_{FE}}. \quad (16.8)$$

Continuous Gaussian wideband interference is interesting because it is dense in both frequency and time and its amplitude distribution is shaped like that of receiver thermal noise. Thus, from the perspective of an adversarial jammer, wideband Gaussian interference is a conservative strategy: Although it demands significant power, it affords receivers in the target area no more effective interference mitigation techniques than those commonly applied for weak GNSS signal tracking.

16.2.2 Narrowband Interference

Suppose $r_I(t)$ is a narrowband interference signal offset by f_I Hz from the GNSS carrier frequency. As an extreme case, consider perfect tone interference

$$\begin{aligned} r_I(t) &= \sqrt{P_I} \exp(j2\pi f_I t) \\ S_{r_I}(f) &= P_I \delta(f - f_I). \end{aligned}$$

In this case, the power spectrum $S_I(f)$ is simply a scaled and frequency-shifted version of $S_{C_I}(f)$

$$\begin{aligned} S_I(f) &= S_{C_I}(f) * S_{r_I}(f + \hat{f}_D) \\ &= P_I S_{C_I}(f) * \delta(f + \hat{f}_D - f_I) \\ &= P_I S_{C_I}(f + \hat{f}_D - f_I). \end{aligned}$$

Smooth Spectrum Approximation

As a first approximation, let $S_{C_I}(f)$ be any smooth (no spectral lines) function with an equivalent rectangular bandwidth of $W_C > 2|f_I|$. Then, interference power P_I/L_C passes into the correlation products, where $L_C = W_C/W_a$ is termed the spread-spectrum processing gain. In this approximation, $I_0 = P_I/W_C$, so that

$$\text{CINR} = \frac{P_S}{N_0 + P_I/W_C}.$$

For a large jamming-to-signal power ratio P_I/P_S , N_0 becomes negligible compared with P_I/W_C , in which case CINR can be approximated as

$$\text{CINR} = 10 \log_{10}(W_C) - 10 \log_{10} \left(\frac{P_I}{P_S} \right) \text{ dB Hz}.$$

For example, if $W_C = 1$ MHz, then a tone interference source with a jamming-to-signal power ratio of $P_I/P_S = 25$ dB would result in a CINR of approximately $60 - 25 = 35$ dB Hz.

Moving toward a more accurate analysis of tone interference, consider now the actual shape of $S_{C_I}(f)$ while retaining the assumption of smoothness (no spectral lines). In particular, suppose that $S_{C_I}(f) = T_C \text{sinc}^2(fT_C)$, which would be the case for a local replica matched to a random binary spreading code $C(t)$ with chip interval T_C . Then, for tone interference with power P_I it follows that

$$\begin{aligned} S_I(f) &= P_I S_{C_I}(f) * \delta(f + \hat{f}_D - f_I) \\ &= P_I T_C \text{sinc}^2[(f + \hat{f}_D - f_I)T_C]. \end{aligned}$$

From this expression, it is clear that the tone interference will minimize CINR (by maximizing $I_0 \equiv S_I(0)$) when $f_I = \hat{f}_D$. In other words, under the smooth spectrum approximation with $S_{C_I}(f) = T_C \text{sinc}^2(fT_C)$, the greatest degradation to CINR occurs when the tone is aligned with the Doppler-shifted carrier frequency of the desired signal.

One can apply a similar analysis to modern GNSS signals with binary offset carrier (BOC) spreading code modulation. In this case, the worst-case tone interference occurs when f_I coincides with the Doppler-shifted peak of one of the offset side lobes. However, due to the additional spreading afforded by BOC-type signals, the resulting interference is, in general, less severe than for a sinc^2 -type waveform with equivalent T_C [16.21].

Effect of Spectral Lines

The smooth-spectrum approximation is appropriate for pseudorandom spreading codes $C(t)$ with a long code repetition period, such as the encrypted legacy military GPS spreading codes, for which the period is not publicly known but surely exceeds one week [16.22], and for the GPS L2CL code, which has a period of 1.5 s [16.23]. For short-period pseudorandom codes, however, the approximation is not appropriate because interference can be narrower than the spacing between spectral lines. Assume that $C(t)$ is a repeating code with period $T_p = T_C N_p$, where $N_p \in \mathbb{N}$ is the number of chips per code period. As a periodic function, $C(t)$ can be

decomposed as a Fourier series, which means that its power spectrum $S_C(f)$ is expressible as a weighted sum of Dirac delta functions

$$S_C(f) = \sum_{i=-\infty}^{\infty} c_i \delta(f - i\Delta f_p), \quad i \in \mathbb{Z} \quad (16.9)$$

with constraint

$$\sum_{i=-\infty}^{\infty} c_i = 1$$

and spectral line spacing $\Delta f_p = 1/T_p$. Assuming a matched local code replica [$C_I(t) = C(t)$], Fig. 16.4 shows the spectral line structure of $S_{C_I}(f)$ for an example GPS L1 C/A code.

For tone interference $S_{r_1}(f) = P_1 \delta(f - f_1)$, $S_I(f)$ is simply a scaled and shifted version of $S_{C_I}(f)$

$$\begin{aligned} S_I(f) &= S_{C_I}(f) * S_{r_1}(f + \hat{f}_D) \\ &= P_1 S_{C_I}(f - f_1 + \hat{f}_D). \end{aligned} \quad (16.10)$$

Interestingly, if none of the tones in the comb of spectral lines that constitute $S_I(f)$ falls within the passband of the accumulate-and-dump filter $H_a(f)$, then the tone

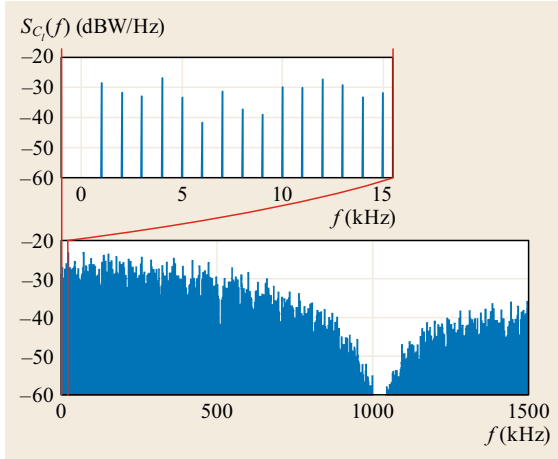


Fig. 16.4 Power spectrum $S_{C_I}(f)$ corresponding to the GPS L1 C/A code replica for pseudo-random number sequence (PRN) 31. The units of $S_{C_I}(f)$ assume that the power of $C_I(t)$ is normalized to 1 W. Because $S_{C_I}(f) = S_{C_I}(-f)$, only positive frequencies are shown. Bottom panel: The interval $0 \leq f \leq 1500$ kHz showing the code's approximate $T_C \text{sinc}^2(f T_C)$ spectral envelope. Top panel: Expanded view of the first 15 kHz, showing distinct spectral lines with irregular weighting spaced at $\Delta f_p = 1/T_p = 1$ kHz

interference will have a negligible effect on the accumulation products. This can be quantified probabilistically as follows. If the frequency offset f_1 is modeled as a random variable uniformly distributed over a range wider than Δf_p , then the probability that one of the spectral lines in $S_I(f)$ will fall within the noise-equivalent bandwidth W_a of the accumulate-and-dump filter is

$$P_X = [\text{mod}(|f_1|, \Delta f_p) \leq W_a] = \frac{W_a}{\Delta f_p}.$$

For N_s signals tracked, each with independent random f_D , the probability of significant interference in any tracking channel rises to

$$P_{X_T} = 1 - (1 - P_X)^{N_s}.$$

By way of example, for GPS L1 C/A-code tracking with $T_a = 20$ ms and $N_s = 10$, $P_X = 0.05$ for each tracking channel and $P_{X_T} = 0.4$ for the ensemble.

From (16.9) and (16.10), it is evident that tone interference is most damaging when f_1 is aligned with the Doppler-shifted spectral line having the largest weighting coefficient c_i . For example, for the spectrum shown in Fig. 16.4, the largest c_i , located at ± 72 kHz, is 23 dB below the total power in $S_{C_I}(f)$. Therefore, when targeting this signal, a tone interferer with power P_1 would be attenuated by at least 23 dB before passing into the accumulate-and-dump filter. (Interestingly, tone interference targeting a C/A signal at exactly the Doppler-shifted L1 carrier frequency is ineffective because the balanced C/A Gold codes, which have only one more 1 than 0, produce a nearly insignificant -60.2 dB line component at zero offset.) In general, the largest spectral line components among all GPS L1 C/A Gold codes attenuate tone interference by only 18.3 dB [16.24]. By way of comparison, a perfectly random code sequence with the same chip interval ($T_C \approx 1 \mu\text{s}$) would attenuate the interferer by at least 60 dB.

In general, one can say that spectral lines in $S_{C_I}(f)$ have two contrary effects on tone interference: (1) line sparsity reduces the probability that interference will have a significant effect – most likely the interference will fall harmlessly between the lines, but (2) in the event that tone interference does coincide with a powerful line component, the interference effect is severe.

Of course, pure tone interference is only a convenient fiction; all interference encountered in practice will have a nonzero spectral width. Convolving an arbitrary $S_{r_1}(f)$ with an $S_{C_I}(f)$ of the form in (16.9) results in an interference spectrum of the form

$$\begin{aligned} S_I(f) &= S_{C_I}(f) * S_{r_1}(f + \hat{f}_D) \\ &= P_1 \sum_{i=-\infty}^{\infty} c_i S_{r_1}(f - \Delta f_p + \hat{f}_D). \end{aligned} \quad (16.11)$$

Thus, each tine in the comb now assumes the shape of $S_n(f)$. For interference that is narrow with respect to Δf_p , each tine remains distinct from its neighbors and is weighted according to the corresponding c_i ; as the interference widens, the tines blend together and the spectrum flattens.

16.2.3 Matched-Spectrum Interference

An inescapable property of multiaccess spread-spectrum systems such as GNSS is that, from the perspective of a receiver channel tracking a particular GNSS signal (a unique combination of spreading code and center frequency), all other signals at the same frequency act as interference. Moreover, many of these interfering signals will have a power spectrum that is closely matched with that of the desired signal. This matched-spectrum interference is a particularly potent nuisance because it allocates power, as a function of frequency, in exact proportion to the weighting that the receiver applies with its local replica in attempting to track the desired signal. Thus, the most powerful spectral lines – the most important contributors to the total received GNSS signal power – are affected by the greatest amount of noise. In recognition of this, adversarial interferers often adopt matched-spectrum interference as their waveform of choice. In the case of nonmalicious intrasystem (e.g., within GPS) or intersystem (e.g., between GPS and Galileo) interference, the competing waveforms are by design weak and approximately power-matched so that the interference is small compared to the ever-present thermal noise, though not entirely insignificant – especially with the proliferation of GNSS satellites.

When matched-spectrum interference originates from GNSS satellites, it is termed multiaccess interference. As an illustration of the effects of such interference, consider a pseudorandom binary spreading code whose power density under a smooth-spectrum approximation is

$$S_C(f) = P_C T_C \text{sinc}^2(f T_C),$$

where P_C is the received signal power and T_C is the spreading code chip interval. This model applies, for example, to the spreading codes of GPS L1 C/A and P(Y), L2 C and P(Y), and L5 I and Q. Assume, for simplicity, that the receiver's power-normalized code replica is perfectly matched to the incoming code so that $S_C(f) = P_C S_{C_i}(f)$ (i.e., band-limiting effects in the RF front end are ignored).

Treating $S_C(f)$ as an interference spectrum and assuming \hat{f}_D is negligible compared to the bandwidth of

$S_C(f)$, we have

$$\begin{aligned} S_I(f) &= S_{C_i}(f) * S_n(f) \\ &= P_C S_C(f) * S_C(f) \\ &= P_C \int_{-\infty}^{\infty} S_C(f - \nu) S_C(\nu) d\nu \\ &= P_C \int_{-\infty}^{\infty} S_C(\nu - f) S_C(\nu) d\nu, \end{aligned}$$

where the last equality follows from $S_C(f) = S_C(-f)$. Hence,

$$\begin{aligned} I_0 \equiv S_I(0) &= P_C \int_{-\infty}^{\infty} S_C^2(\nu) d\nu \\ &= P_C \int_{-\infty}^{\infty} [T_C \text{sinc}^2(\nu T_C)]^2 d\nu \end{aligned}$$

which, by the change of variables $q = \nu T_C$, becomes

$$I_0 = P_C T_C \int_{-\infty}^{\infty} \text{sinc}^4(q) dq = \left(\frac{2}{3}\right) P_C T_C.$$

Thus, the effect of a single multiaccess interference signal with received power P_C is to raise the effective thermal noise density from N_0 to

$$N_{0,\text{eff}} = N_0 + \left(\frac{2}{3}\right) P_C T_C.$$

The significance of multiaccess interference is measured with respect to N_0 . Suppose there are M multiaccess signals whose average received power is \bar{P}_C . Then, from the perspective of a single desired signal, the multiaccess power density becomes equivalent to N_0 when

$$\left(\frac{2}{3}\right) \bar{P}_C T_C (M - 1) = N_0.$$

Thus, to ensure that multiaccess density does not exceed N_0 requires

$$M \leq 1 + \frac{3/2}{(\bar{P}_C/N_0) T_C}.$$

Figure 16.5 shows this bound for $T_C = 1 \mu\text{s}$, which applies to GPS L1 and L2C, and for $T_C = 0.1 \mu\text{s}$, which

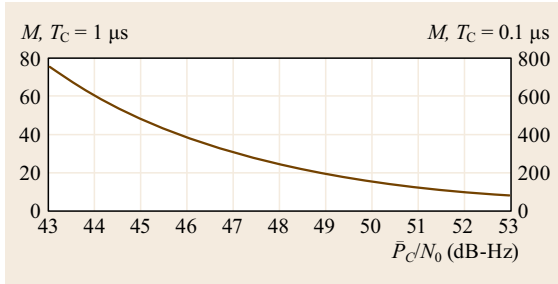


Fig. 16.5 Maximum number of simultaneously received multiaccess GNSS signals with power spectrum $S_C(f) = P_C T_C \text{sinc}^2(fT_C)$ such that $I_0 \leq N_0$, as a function of \bar{P}_C/N_0 , where \bar{P}_C is the average power of the $M-1$ multiaccess interferers. The left- and right-hand scales correspond, respectively, to $T_C = 1 \mu s$ and $T_C = 0.1 \mu s$

applies to GPS L5 I and Q. Assuming that, for the average user, the number of received signals M is approximately one-fourth of the total number of orbiting

GNSS satellites and that $\bar{P}_C = 47 \text{ dB Hz}$, and assuming all satellites broadcast only the GPS L1 C/A signal, the multiaccess interference density exceeds N_0 when the constellation size grows beyond 124 satellites.

It is worth noting that, although a 3 dB rise in the effective thermal noise floor (from N_0 to $N_0 + I_0 = 2N_0$) is significant, most GNSS users would gladly trade this degradation for the vastly improved dilution of precision and reduced convergence times for carrier-phase differential GNSS (CDGNSS) positioning and precise point positioning (PPP) that a larger multi-GNSS constellation would afford.

Finally, observe that, from the perspective of an adversarial interferer, matched-spectrum interference is the most efficient use of transmit power among all interference waveforms. For example, in the case of a local replica with density $S_C(f) = T_C \text{sinc}^2(fT_C)$, it can be shown that for a fixed interference power P_I , the interference density assumes its maximum value $I_0 = (2/3)P_I T_C$ when $S_{r1}(f) = P_I T_C \text{sinc}^2(fT_C)$.

16.3 Quantization Effects

The effect of signal quantization on interference depends less on the bandwidth of the interference – whether wideband or narrowband – than on its amplitude distribution. The salient result in this regard is as follows: For white, Gaussian-distributed interference, the quantizer's output SNR is always degraded relative to its input SNR, whereas for constant-amplitude interference (e.g., a swept tone), the quantizer output SNR can actually exceed its input SNR. In any case, an optimal quantization strategy seeks to minimize the SNR degradation through the quantizer.

16.3.1 One-Bit Quantization

If the discrete samples entering a one-bit (two-level) quantizer are Gaussian distributed and uncorrelated, then the SNR is degraded by a factor $2/\pi$ or -1.96 dB [16.25]. Designers of low-cost GNSS receivers often view this modest loss as a small price to pay for a one-bit quantizer's economy of implementation and low power consumption, which explains the popularity of one-bit quantization in consumer devices.

However, one-bit quantization performs poorly in the presence of strong tone interference [16.24]. To understand why, consider a simple case in which thermal noise is absent and a pure tone interference signal is received phase coherently (in-phase) with the carrier of a desired biphas-modulated GNSS signal. In this case, it is clear that, if the interference amplitude α

is greater than the GNSS signal amplitude, then the interference completely suppresses the GNSS signal in one-bit quantization because the signal's noise-free biphas transitions are dominated at every sampling instant by the coherent interference.

In the presence of thermal noise, the desired GNSS signal is no longer completely suppressed by coherent tone interference, but the quantizer SNR degradation remains severe whenever $\alpha > \sigma$, where σ is the thermal noise standard deviation. Note that if the tone interference is out of phase by some angle θ , then its effective amplitude becomes $\alpha \cos \theta$. Thus, if θ is slowly varying and $\alpha > \sigma$, then the GNSS signal is periodically suppressed. When θ varies rapidly compared to the reciprocal integration time $1/T_a$, as with tone interference significantly offset from the desired GNSS signal carrier frequency – or, more generally, with any constant-amplitude interference – SNR degradation is less severe than in the case of coherent tone interference but still increases rapidly with increasing $\alpha > \sigma$.

It follows from these observations that one-bit quantization is a serious design flaw for receivers meant to operate in the presence of strong constant-amplitude interference.

16.3.2 Multibit Quantization

Multibit quantization is preferable to one-bit quantization when constant-amplitude interference may be

present. Not only can multibit quantization prevent total suppression of the desired GNSS signal, but, with properly chosen quantization levels, it can substantially suppress constant-amplitude interference.

Two-bit (four-level) quantization is an especially attractive option for GNSS receivers because it is simple to implement and amenable to low-power processing yet yields significantly less SNR degradation than one-bit quantization in wideband Gaussian noise (0.55 dB versus 1.96 dB [16.24, 26, 27]). The two-bit quantization function $q_2(x)$ is graphically shown in Fig. 16.6. For uncorrelated zero-mean Gaussian noise with standard deviation σ , both the minimum mean-square-error distortion criterion [16.28] and the minimum SNR degradation criterion [16.26] (in the limit of low SNR) are optimized when the magnitude threshold is chosen as $L = 0.98\sigma$ and the ratio of the quantization levels is approximately $a_2/a_1 = 3.3$. This remains true whether the noise is thermal in origin (i. e., proportional to the receiver system temperature) or is a combination of thermal noise and ambient interference, so long as the combined noise-plus-interference amplitude distribution remains Gaussian and sample-wise uncorrelated. Implementation of this quantization strategy within a GNSS receiver is typically realized by setting $a_1 = 1$, $a_2 = 3$ and adjusting the automatic gain control (AGC) so that $|q_2(x)| = a_2$ with probability 0.33.

When significant non-Gaussian interference is present in the received analog signal, the probability distribution $p(x)$ of the input to the quantizer is no longer approximately Gaussian and the above values for a_1 , a_2 , and L become suboptimal. If $p(x)$ is known, then new mean-square-distortion-minimizing values can be calculated numerically as described in [16.28]. For the special case of unity-amplitude tone interference with a phase that varies rapidly relative to $1/T_a$, and in

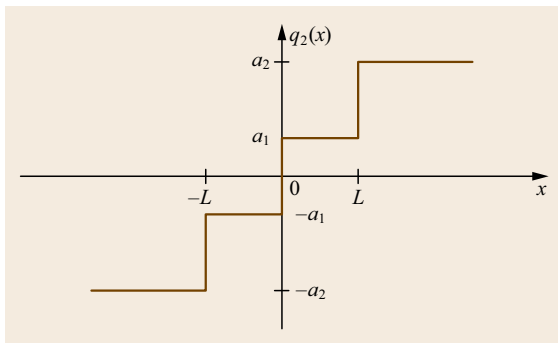


Fig. 16.6 Quantization function $q(x)$ for two-bit (four-level) quantization, showing the magnitude threshold L and the quantization levels $\{-a_2, -a_1, a_1, a_2\}$

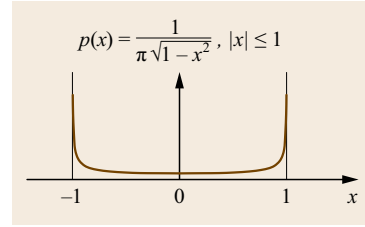


Fig. 16.7 Probability distribution of the quantizer input x for unity amplitude tone interference in the limit of low SNR

the limit of low SNR, $p(x)$ assumes the shape shown in Fig. 16.7. In this case, it can be shown numerically that the mean-square distortion is minimized when $L = 0.573$ and $a_2/a_1 = 2.89$. But, importantly, and in contrast to the Gaussian noise-plus-interference case, these distortion-minimizing values do not also minimize SNR degradation. Instead, for spread-spectrum signals with large processing gain (such as GNSS signals), SNR degradation is minimized as L approaches the upper limit of $p(x)$ [16.26]. The key insight is that, for this choice of L , the quantizer maximizes the number of captured code transitions, as illustrated in Fig. 16.8.

More generally, a properly configured multibit quantizer exhibits *negative* SNR degradation (i. e., there is a positive *conversion gain*) when the incoming interference has a fixed amplitude (e.g., a swept tone). This result holds even when the interference is a combination of fixed-amplitude and Gaussian interference, so long as the fixed-amplitude interference dominates [16.29]. This contrasts with Gaussian interference, for which a two-bit quantizer's output SNR is always degraded by at least 0.55 dB relative to its input SNR.

Within a GNSS receiver, adaptive two-bit quantization for suppression of constant amplitude interference can be implemented as follows. When significant constant-amplitude interference is detected, the adaptive quantizer raises the threshold L from the Gaussian-noise-optimized value for L (approximately

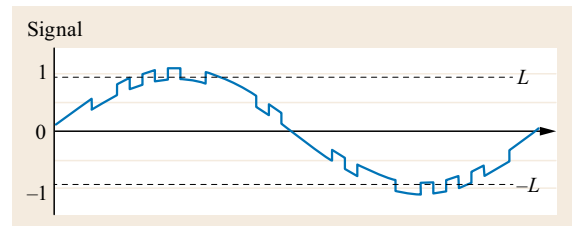


Fig. 16.8 Example threshold value L for two-bit quantization of a binary spread-spectrum signal in the presence of strong unity-amplitude tone interference. As the signal-to-interference power ratio decreases from the -20 dB ratio shown, the curve's distribution approaches that of Fig. 16.7, and the optimal value of L approaches 1

$L = \sigma$) to a new value that places L near the edge of the $p(x)$ distribution (equivalently, the AGC can lower its gain until this condition is reached). The optimal value of L depends on the relative strengths of the GNSS signal, the constant-amplitude interference, and the Gaussian noise and interference. Figure 16.9 shows the quantizer conversion gain for several example scenarios with different relative signal, noise, and interference strengths. A simple suboptimal approach sets L so that $|q(x)| = a_2$ with a predetermined probability (e.g., 10%); in an alternative, higher-performance approach, a feedback signal from the GNSS receiver's baseband processor adjusts L to maximize the average C/N_0 of the tracked GNSS signals. Note that as the constant-amplitude interference power increases relative to the Gaussian interference, the quantizer can more effectively suppress the former, but its performance becomes more sensitive to choice of L . For best performance, the ratio a_2/a_1 should also be adjusted upward from its Gaussian-adapted setpoint (approximately $a_2/a_1 = 3$), but this is less important than adjusting L . An example of adaptive multibit quantization implementation can be found in [16.14, Fig. 6.1].

Three-bit (8-level) and higher quantization bring further reduction of SNR degradation for all interference and noise types, but the marginal improvement above two-bit quantization is modest and decreases rapidly with additional bits. In uncorrelated Gaussian noise and interference, the SNR degradation through a three-bit quantizer is 0.272 dB (versus 0.55 dB for a two-bit quantizer) [16.27]. Details on three-bit quantizer performance can be found in [16.24].

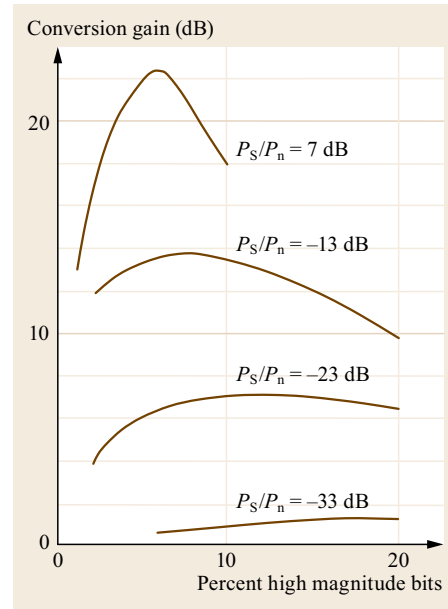


Fig. 16.9 Two-bit quantizer conversion gain (ratio of quantizer output SNR to input SNR) for a scenario in which the incoming spread-spectrum signal is corrupted by both Gaussian noise (or interference) and constant-amplitude interference, as a function of the percentage of high magnitude bits (percentage of samples for which $|q(x)| = a_2$). The different curves correspond to different values of the signal power to Gaussian noise (or interference) ratio P_S/P_n . For all curves, the ratio of the signal power to the constant-amplitude interference is $P_S/P_{ca} = -40$ dB, and $a_2/a_1 = 8$ (after [16.29], courtesy of the Institute of Electrical and Electronics Engineers (IEEE))

16.4 Specific Interference Waveforms and Sources

16.4.1 Solar Radio Bursts

Solar radio bursts (SRBs) are intense outbursts of radio emissions from the Sun, with spectral power ranging from HF to above the L band. They are typically associated with solar flares, which are caused by the acceleration of electrons in the solar atmosphere and whose rate of occurrence follows the 11 yr sunspot cycle [16.30, 31]. SRBs' jamming effect on radio equipment was first noted during World War II when strong SRBs jammed British anti-aircraft radar on many occasions [16.32]. SRBs can cause greater than 10 dB fades in a GNSS signal's C/N_0 [16.33, 34].

Given their broad-spectrum power distribution, SRBs are typically modeled as contributing to a receiver's thermal noise $n(t)$. In particular, they raise

a GNSS receiver's antenna temperature T_A , which is related to the receiver's noise density N_0 by

$$N_0 = k_B(T_R + T_A),$$

where k_B is Boltzmann's constant and T_R and T_A are respectively the receiver and antenna noises in degrees Kelvin. T_R is the equivalent temperature of noise sources internal to the receiver, primarily those in the first-stage low-noise amplifier (LNA). T_A is the temperature equivalent of noise impinging on the antenna, including radiation from the warm Earth, cosmic noise, and solar radio noise. T_A varies with antenna motion (as more or less warm Earth radiation is visible), antenna blockage (e.g., an increase in T_A due snow accumulation [16.35]), and variable solar radiation. Note that

these are difficult or impossible for a stand-alone (non-networked) GNSS receiver to predict. Of these, solar radiation is least site-specific: All GNSS receivers in view of the Sun are similarly affected.

To judge the impact of SRBs on GNSS receivers, it is instructive to examine the rate of occurrence of those SRBs that significantly increase a receiver's P_T . Such events not only reduce C/N_0 but also lead to false alarms in received power monitoring, a technique whereby intentional interference is detected based solely on P_T (discussed further in Sect. 16.6.2). Table 16.2 shows the SRB occurrence rate for three different levels of increased P_T . Let $P_T/P_{T,\text{nom}}$ be the ratio of received power in the presence of a SRB to nominal received power. Assume that non-SRB interference is negligible so that $P_I = 0$, leaving $P_T = P_S + P_n$, where

$$P_n = W_{\text{FE}}N_0 = W_{\text{FE}}k_B(T_R + T_A).$$

Let the antenna temperature be $T_A = T_{A0} + T_{As}$, where T_{A0} is a nominal value for T_A and T_{As} is the increase in T_A due to solar radiation.

Table 16.2 is interpreted as follows. Each value of $P_T/P_{T,\text{nom}}$ can be related to a value of T_{As} by

$$\frac{P_T}{P_{T,\text{nom}}} = \frac{P_S + k_B B(T_R + T_{A0} + T_{As})}{P_S + k_B B(T_R + T_{A0})}$$

assuming the following reasonable parameter values: $P_S = -146$ dBW, $W_{\text{FE}} = 2$ MHz, $T_R = 188$ K, and $T_{A0} = 100$ K. Each T_{As} , in turn, is related to a change in C/N_0 by

$$\Delta C/N_0 = \frac{T_R + T_{A0}}{T_R + T_{A0} + T_{As}}$$

and to a solar flux density S_1 by

$$S_1(\text{SFU}) = \frac{2k_B T_{As}}{A_e 10^{-22}},$$

where the effective antenna area is taken to be $A_e = 7.23 \cdot 10^{-3} \text{ m}^2$, which is a good approximation for a single-element GNSS antenna, and where the additional factor of 2 in the numerator reflects the assumption that only half the total-polarization solar radiation

contributes to T_{As} through a GNSS antenna, which is designed to receive right-hand circularly polarized signals [16.34]. The factor 10^{-22} converts $\text{W/m}^2/\text{Hz}$ to solar flux units (SFU). The resulting S_1 values listed in Table 16.2 are those above which P_T would increase by the amount shown. As a final step, the model $N(S > S_1, \nu_1, \nu_2)$ from [16.36] is invoked (with the correction factor C_{geo}) to approximate the total number of bursts exceeding S_1 in the frequency range ($\nu_1 = 1$ GHz, $\nu_2 = 1.7$ GHz) over a 40 yr historical period. This is used to estimate T_e , the time between triggering events, for solar maximum years and for all years.

Table 16.2 reveals that solar radio bursts causing a degradation in C/N_0 of 1.9 dB or greater are rare, occurring approximately once per month during solar maximum. Truly intense SRBs causing 10 dB or more of degradation and interrupting signal tracking, as in the 2006 storm [16.33], are extremely rare. Nonetheless, SRBs can be problematic for signal authentication techniques based solely on P_T , as will be discussed in Sect. 16.6.2.

16.4.2 Scintillation

A transionospheric radio wave can exhibit temporal fluctuations in phase and intensity caused by electron density irregularities along its propagation path, a phenomenon called scintillation, or fading. At GNSS frequencies (L band), strong scintillation is manifest in deep power fades (> 15 dB) that are often associated with rapid phase changes. Such vigorous signal dynamics stress a receiver's carrier tracking loop and, as their severity increases, lead to navigation bit errors, cycle slipping, and complete loss of carrier lock [16.37, 38].

Signal refraction, caused by large-scale irregularities, results in low-frequency variations in group delay (measured by the code phase, or pseudorange, observable) and carrier phase. Signal diffraction, caused by smaller-scale (approximately 400 m) irregularities, scatters L-band signals so that the radio waves reach terrestrial receivers through multiple paths. Interaction between signals from multiple directions occurs at the carrier-phase level, yielding constructive and destructive interference patterns that produce variations in both the phase and amplitude of received signals.

It may at first seem out of place to treat ionospheric scintillation as interference, but the mutual interference caused by diffraction can challenge signal tracking as much as intermittent jamming, and diffractive interference shares characteristics with structured interference such as GNSS spoofing. The same argument can be made for nonionospheric multipath effects – those due to signal reflections – but these are treated separately in Chap. 15. Chapter 39 also treats scintillation, but

Table 16.2 Time between threshold-exceeding solar radio burst events for various values of the ratio $P_T/P_{T,\text{nom}}$

Threshold values				T_e (days)	
$P_T/P_{T,\text{nom}}$ (dB)	T_{As} (K)	$\Delta C/N_0$ (dB)	S_1 (SFU)	Solar max.	All years
0.44	40.9	−0.6	1560	9.2	22.0
0.93	91.3	−1.2	3488	17.3	42.9
1.5	157.7	−1.9	6022	26.5	67.4

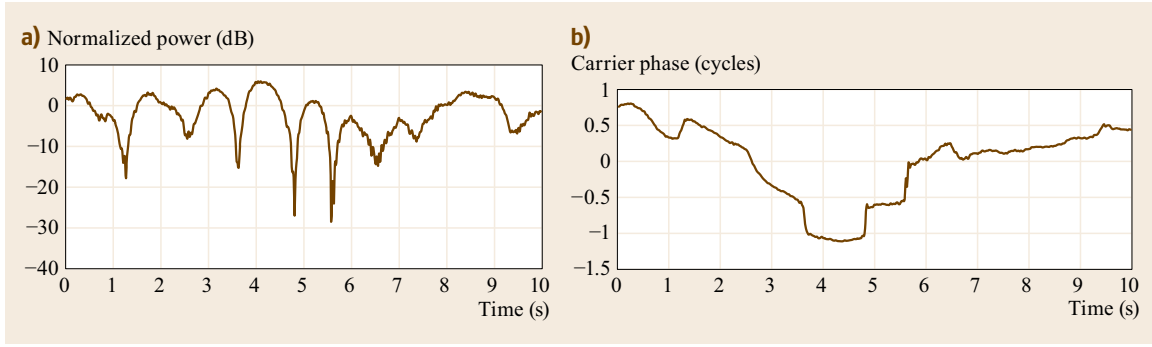


Fig. 16.10a,b Normalized signal power (a) and carrier phase (b) time histories from a record of GPS L₁ data with $S_4 \approx 0.9$ (after [16.37], courtesy of IEEE)

with an eye to phenomenology rather than receiver effects.

Severe L-band scintillation is both infrequent and geographically confined. The type known as equatorial scintillation, or equatorial spread F, generally occurs between local sunset and 2400 local time in the region extending $\pm 15^\circ$ about the magnetic equator [16.39]. Another common type of scintillation occurs at high latitudes [16.40]. Significant effects have also been noted in the mid-latitude region, but they occur infrequently [16.41]. This section concentrates on equatorial scintillation because it is the most interference-like, making signals particularly difficult to track.

The severity of scintillation can be succinctly characterized by two parameters, the scintillation index, S_4 , and the decorrelation time τ_0 [16.42]. S_4 measures the intensity of scintillation, and is defined by

$$S_4^2 = \frac{\langle I^2 \rangle - \langle I \rangle^2}{\langle I \rangle^2},$$

where $I = \alpha^2$ is signal intensity, α being the signal amplitude, and $\langle \cdot \rangle$ denotes time average. The scintillation decorrelation time $\tau_0 > 0$ is a measure of the rapidity of scintillation. A small τ_0 (e.g., < 0.5 s) implies a scintillating channel that changes rapidly with time.

A short sample from the scintillation library introduced in [16.37] is presented in Fig. 16.10. The sample manifests strong scintillation, with $S_4 \approx 0.9$. The most striking features of the plot are the deep power fades that occur simultaneously with abrupt, approximately half-cycle phase changes whose sense (downgoing or upgoing) appears random. Such fades appear to be a universal feature of strong equatorial scintillation, and they are the primary cause of phase unlock for PLLs tracking strongly scintillating signals.

PLLs are affected by scintillation in two related ways: (1) increased phase error variance and (2) phase unlock.

Phase Error Variance

The phase error variance models given in Sect. 16.1.4 assume that all phase errors are due to constant-intensity white measurement noise. Furthermore, (16.5) and (16.6) assume PLL linearity. These assumptions are violated during severe scintillation: Amplitude fading causes variations in the loop SNR, phase changes are time correlated, and, when attempting to track through the large, rapid phase changes associated with deep fading, the PLL cannot be expected to operate in its linear regime. For these reasons, calculating the phase error variance for a PLL tracking through strong scintillation is not straightforward [16.38]. Figure 16.11 shows how

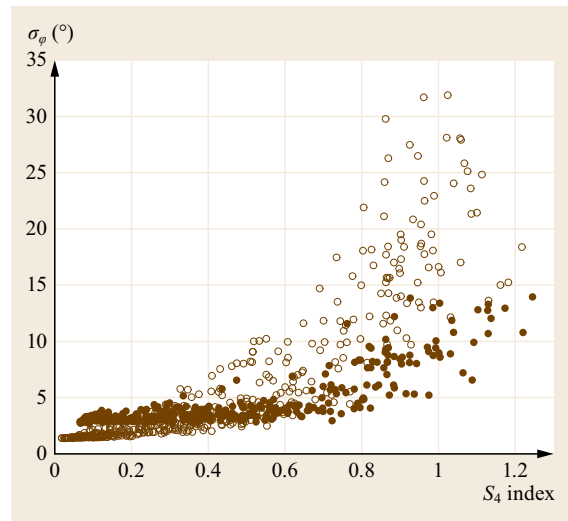


Fig. 16.11 Standard deviation of PLL phase error modulo π for a decision-directed arctangent phase discriminator over 30 s test records versus S_4 for ultra-high frequency (UHF) signals at $C/N_0 = 43$ dB Hz (open circles) and for GPS L₁ signals within $40 < C/N_0 < 44$ dB Hz with mean $C/N_0 = 43$ dB Hz (filled circles) (after [16.38], courtesy of IEEE)

σ_φ , the standard deviation of the phase measurement error modulo π , increases with increasing S_4 , a dependence that is both due to the fade-induced reductions in loop SNR and to phase scintillation with frequency components that exceed the PLL's bandwidth. The large values of σ_φ at high S_4 contribute to the degradation of carrier-phase-dependent GNSS systems during strong scintillation.

Phase Unlock

The general term *phase unlock* refers to single or successive cycle slips. Phase and amplitude scintillation cause cycle slipping by either deep rapid fading or prolonged fading. In the limit as the fade depth increases, the accompanying abrupt, nearly π -rad phase transition looks like bi-phase data modulation, to which a squaring-loop PLL is insensitive by design. Hence, the PLL detects no phase shift and a half-cycle slip occurs. In marginal cases, where the PLL might be capable of distinguishing a scintillation-induced phase transition from a data-bit-induced phase transition, the sudden drop in loop SNR increases the likelihood of a cycle slip. In short, simultaneous power fades and abrupt phase changes are a particularly challenging combination.

Prolonged amplitude fading is the second mechanism by which scintillation causes cycle slipping. This phenomenon may be considered a special case of fading in which the fading time scale is elongated so that the amplitude fade is accompanied by phase dynamics that are slow compared to a typical 10 Hz PLL noise bandwidth. In this case, broadband measurement noise dominates and (16.7) applies. Cycle slips occur rarely by this mechanism.

Figure 16.12 presents results in terms of cycle slip rate on the left vertical axis, and, for convenience, in terms of the mean time between slips, T_s , on the right vertical axis. As would be expected, a general increase in the rate of cycle slips accompanies increasing S_4 . The lack of cycle slips below $S_4 \approx 0.4$ suggests that, whatever its other characteristics (e.g., τ_0), scintillation with $S_4 \lesssim 0.4$ can be considered benign.

16.4.3 Unintentional Interference

Spectral surveys of the GNSS bands reveal that in rural areas the bands are largely free of interference, but in urban areas they are often corrupted by intermittent interference sources [16.43]. Most of these interference events are unintentional. Similarly, radio frequency interference (RFI) can disturb signal tracking when a GNSS receiver's antenna is packaged closely to other electronic equipment, as on a small satellite. Fol-

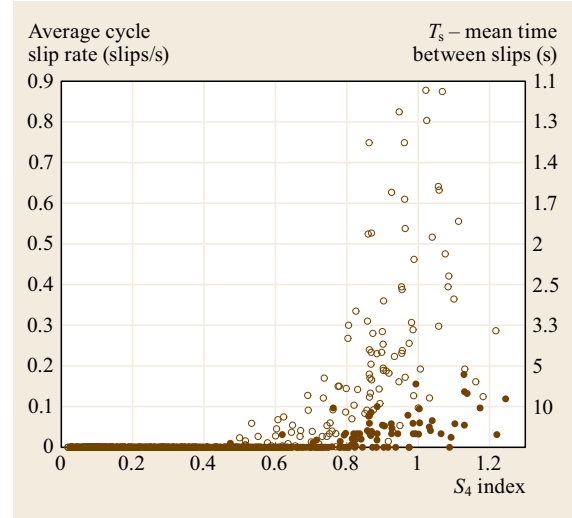


Fig. 16.12 Average cycle slip rate for the decision-directed arctangent phase discriminator over 30 s test records versus S_4 for UHF signals at $C/N_0 = 43$ dB Hz (open circles) and for GPS L1 signals within $40 < C/N_0 < 44$ dB Hz with mean $C/N_0 = 43$ dB Hz (filled circles). The right vertical axis expresses the cycle slip rate in terms of T_s (after [16.38], courtesy of IEEE)

lowing are some examples of unintentional interference sources.

Harmonics

Nonlinearity in any one of several stages involved in RF transmission generates power not only at the intended transmission frequency but also at integer multiples, or harmonics, of that frequency. For example, UHF television signals with carrier frequencies near 525 MHz are notorious for injecting third-harmonic power into the GNSS L1 band [16.44, 45].

When broadcast transmitters are powerful, as with television transmitters, a harmonic near the GNSS bands can substantially degrade GNSS tracking performance. If a harmonic lies within a GNSS band of interest, then it cannot be attenuated by standard RF filters designed to isolate the GNSS signals. If powerful enough, the interfering harmonic will drive a GNSS receiver's dominant LNA into its nonlinear regime, causing a loss of sensitivity and leaving spurious tones across the target GNSS band [16.45].

DME/TACAN

The GPS L5 band and the Galileo E5a and E5b bands are situated in an ARNS band also allocated to distance measuring equipment (DME) and Tactical Air Navigation (TACAN) systems whose strong pulsed emissions act to significantly degrade GNSS

tracking [16.1]. DME/TACAN systems, which operate between 960 and 1215 MHz, produce emissions that are sparse in both the time and frequency domains. Pulses are transmitted in pairs 12 μ s apart, with each pulse lasting 3.5 μ s. The maximum practical transmission rate is 2700 pulse pairs per second, which means that interference from a single DME/TACAN transmitter is limited to less than 2% of a 1 s time interval. In the frequency domain, a single DME/TACAN signal occupies only 100 kHz, with channels spaced by 1 MHz. Thus, the total time-frequency occupancy of a single DME/TACAN transmitter in a 10 MHz band is only 0.02%. Such sparsity permits mitigation techniques that render DME/TACAN interference harmless even when GNSS receivers are airborne over so-called hot spots having a high density of DME transmitters [16.1].

Powerful Near-Band Transmissions

The radio spectrum between 700 MHz and 2 GHz, which includes all current GNSS bands, is particularly attractive for the provision of data to mobile units such as smartphones because the wavelengths of signals in this band are short enough that small antennas can be effective yet long enough to penetrate indoors. These desirable properties, coupled with the intense and rising demand for mobile data, portend the eventual placement of powerful transmissions in the radio bands adjacent to GNSS bands.

The 2010–2012 debate over whether to allow powerful terrestrial long term evolution (LTE) signals to be broadcast in the mobile satellite service (MSS) band just below the GNSS L1 band brought to the fore the susceptibility of contemporary GNSS receivers, especially high-precision receivers, to powerful near-band transmissions [16.46]. It was shown, for example, that typical GPS and Galileo receivers tracking signals centered at 1575.42 MHz suffered C/N_0 degradation greater than 3 dB when exposed to communications signals with received power exceeding -80 dBm in the 1545.2–1555.2 band even when the latter were filtered with a high-quality bandpass filter [16.8].

16.5 Spoofing

A GNSS spoofing signal is a type of structured interference that adheres closely enough to a GNSS signal specification so as to appear authentic to an unsuspecting GNSS receiver. Whether intentional, as in a deliberate attempt to manipulate the PVT readout of a target GNSS receiver [16.50, 51], or unintentional, as in an errant GNSS simulator or repeater signal, spoofing

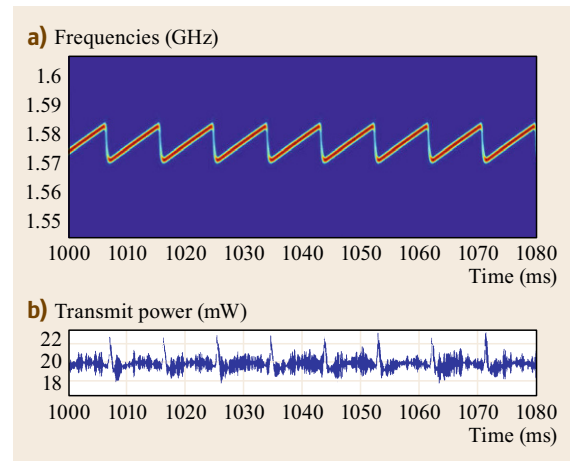


Fig. 16.13a,b Time histories of frequency spectrum (a) and transmit power (b) for a typical chirp-style PPD (after [16.47])

16.4.4 Intentional Interference

Intentional interference, or jamming, has been a staple of navigation warfare since World War II [16.32]. With the emergence of PPDs [16.7] and incidents of nation-scale intentional disruption of civil GNSS [16.48], intentional interference is now also a civil concern.

PPDs are by far the most common source of intentional interference. The PPD user may intend only to jam GNSS tracking devices in his near vicinity (e.g., on his person or vehicle), but in fact such devices can disrupt GNSS signal tracking out to an effective radius of from 100 m to several kilometers [16.47].

Virtually all PPDs transmit a swept tone waveform (chirp) similar to that shown in Fig. 16.13. This waveform can be generated from inexpensive components and is quite effective in rendering GNSS receivers inoperable unless these have been especially designed for jam resistance [16.49]. The frequency sweep period of the 18 units tested in [16.47] ranged from 1 to 27 μ s, with total transmit power in a 20 MHz band centered at L1 ranging from -14 to 28 dBm.

signals similarly affect a GNSS receiver. For convenience of presentation, the following discussion will treat all spoofing as intentional, with the term spoofer referring both to the spoofing device and its operator.

Spoofing was once only a threat to military GNSS receivers and applications, but has now become a more

general concern as civil GNSS spoofing becomes easier and its consequences are more serious. The emergence of low-cost off-the-shelf software-defined radio hardware has significantly reduced the cost and complexity of spoofing. With such hardware, a competent programmer sufficiently familiar with the openly documented GNSS protocols [16.23, 52] can generate realistic civil GNSS signals despite having minimal knowledge of RF electronics. Easier still, low-cost GNSS signal simulators and record-and-replay devices enable even GNSS neophytes to conduct a limited but potent form of spoofing. Against a backdrop of increasing economic dependence on civil GNSS for transportation, communication, finance, and power distribution, the increased accessibility of civil GNSS spoofing raises the risk of attack and the urgency of finding effective antispoofing measures.

Spoofing is different from unstructured interference in two primary respects. First, it can be surreptitious: Neither the target GNSS receiver nor its operator may detect that an attack is underway because the spoofer can seamlessly supplant counterfeit signals for their authentic counterparts. Second, in a spoofing attack, the received interference $r_I(t)$ is statistically correlated with the received authentic signal $r_S(t)$; consequently, the total received power P_T is neither the sum of P_S , P_I , and P_n , as in (16.2), nor does the autocorrelation function of the interference component $I(t)$ decompose, as in (16.4), because the cross-terms do not average to zero. As a result, the analysis of spoofing effects is, in general, more challenging than the analysis of statistically independent interference. To be sure, spoofing effects bear a strong resemblance to multipath effects, but multipath-induced structured interference is accidental, whereas spoofing may involve a strategic attacker who can arbitrarily adjust signal power, code phase, carrier phase, and signal structure for maximum effect.

To generalize the treatment of spoofing in what follows, the authentic signal model will allow for digital modulation that is unpredictable to a would-be spoofer. A modulation sequence that is entirely unpredictable or has unpredictable segments will be termed a security code, and a security-code-bearing GNSS signal will be termed security enhanced [16.53–57]. A nonsecurity-enhanced GNSS signal can be represented by a special case of this model in which the security code is replaced by a sequence of ones.

16.5.1 Generalized Model for Security-Enhanced GNSS Signals

From the perspective of a GNSS receiver, current and proposed security-enhanced GNSS signals can be represented by a simple adaptation of the baseband received signal model introduced in (16.1):

resented by a simple adaptation of the baseband received signal model introduced in (16.1):

$$\begin{aligned} r_S(t) &= \sqrt{P_S} W(t-\tau) D(t-\tau) C(t-\tau) \exp(j\theta(t)) \\ &= \sqrt{P_S} W(t-\tau) X[\tau, \theta(t)]. \end{aligned} \quad (16.12)$$

Compared to (16.1), the novel component here is $W(t)$, which represents a ± 1 valued security code with chip length T_W . For notational simplicity, the product of the authentic signal's navigation data stream $D(t-\tau)$, spreading (ranging) code $C(t-\tau)$ and baseband phasor $\exp(j\theta)$ is abbreviated as $X(\tau, \theta)$ for code phase τ and carrier phase θ . The chip length of the spreading code $C(t)$ is denoted as T_C . For convenience, receiver time t is assumed to be equivalent to true time (e.g., GPS system time).

The security code $W(t)$ is either fully encrypted or contains periodic authentication codes. The defining feature of $W(t)$ is that some or all of its symbols are unpredictable to a would-be spoofer prior to broadcast from a legitimate GNSS source. The unpredictable symbols in $W(t)$ serve two related functions: (1) they enable verification of $W(t)$ as originating from a GNSS Control Segment (standard message authentication), and (2) they increase the complexity of a spoofing attack by forcing the spoofer to either replay a received $W(t)$ or attempt to estimate $W(t)$ on-the-fly. Note that if a GNSS signal is not security enhanced (has no unpredictable modulation), the model in (16.12) still applies, with $W(t) = 1$.

16.5.2 Attacks Against Security-Enhanced GNSS Signals

The unpredictability of the security code $W(t)$ is an obstacle for a would-be spoofer. A simple spoofing technique, such as discussed in [16.58], relies on the known signal structure of the GPS L1 C/A signal and the near-perfect predictability of its navigation data stream. However, if a GNSS signal is security enhanced, then the spoofer of [16.58] cannot perfectly match its counterfeit signals chip-for-chip to the authentic signals.

A spoofer could, of course, ignore the broadcast security codes altogether, filling in dummy values for $W(t)$, but such a scheme is easily detected. In an attack against a GNSS signal modulated by a low-rate security code ($T_W \gg T_C$) (e.g., navigation message authentication (NMA), as proposed in [16.55–57, 59]), the dummy $W(t)$ values would fail the cryptographic validation test. Against a high-rate security code ($T_W \approx T_C$), the dummy $W(t)$ values would yield zero av-

erage power when correlated with the true $W(t)$ sequence [16.53, 59].

Therefore, to be effective while evading detection, a spoofer must attempt to match both the structure and content of the authentic signal. It can do this via one of the following specialized spoofing attacks.

Meaconing

A meaconing, or replay, attack is a specialized spoofing attack in which an entire segment of RF spectrum is captured and replayed [16.60]. If the meaconer employs a single receiving antenna element, then no individual signal is isolated in a meaconing attack. Thus, in this case, a GNSS meaconer cannot arbitrarily manipulate the PVT of a target receiver. Rather, the target receiver will display the position and velocity of the meaconer's receive antenna and a time in arrears of true time. If this antenna is on a dynamic platform, then the meaconer can adjust the position and velocity implied by its signals for greater effect in the attack.

If the meaconer employs multiple antenna elements whose RF signals are individually digitized, then it can isolate individual GNSS signals by pointing a gain enhancement toward each overhead GNSS satellite. For example, a 16 element antenna array could be used to direct a narrow ≈ 12 dB enhancement toward each satellite. By combining the separate digital streams while manipulating the phasing of each stream within the ensemble, a meaconer can dictate the ensemble's implied PVT within a wide range about the true PVT (with the implied timing always in arrears of true time).

For a single GNSS signal corresponding to a particular satellite, the combined meaconed and authentic received signals can be modeled as (16.1) but with $r_S(t)$ as in (16.12) and

$$r_1(t) = \alpha \sqrt{P_S} W(t - \tau_c) X[\tau_c, \theta_c(t)] + n_c(t).$$

Here, $\tau_c > \tau$ and θ_c are the code phase and carrier phase of the counterfeit meaconing signal, respectively, and $n_c(t)$ is the noise introduced by the meaconer's RF front end. The meaconed signal arrives at the target receiver's antenna with a delay $d = \tau_c - \tau > 0$ seconds relative to the authentic signal, an unavoidable consequence of the triangle inequality and the processing delay through the meaconing device. The coefficient α is the meaconed signal's amplitude advantage factor relative to the authentic signal.

High-performance digital signal processing hardware permits a meaconer located close to its intended target to drive the delay d to under a few tens of nanoseconds. In the limit as d approaches zero, the attack becomes a zero-delay meaconing attack in which the meaconed signals are code-phase-aligned with their

authentic counterparts. Such alignment enables a seamless liftoff of the target receiver's tracking loops, following which a meaconer can increase d at a rate that is consistent with the target receiver's clock drift and gradually impose a significant timing delay.

Note that, unless $d \approx 0$, a meaconer with $\alpha \approx 1$ will cause significant variations in the target receiver's PVT estimate: the meaconing signals will act as severe multipath. Thus, if the meaconer cannot ensure $d \approx 0$, it is better off transmitting with an overwhelming amplitude advantage ($\alpha \gg 1$) to quickly stabilize the target's perceived PVT at the meaconer's intended value. Therefore, a meaconer with d a significant fraction of T_C is detectable at $\alpha \approx 1$ due to multipath-like PVT variations and at $\alpha \gg 1$ due to anomalous high received power. Furthermore, if $d > 2T_W$, then the meaconer will be unable to capture a code tracking loop that is locked to an authentic signal for any value of α : The meaconing signal will not be close enough in time to the authentic signal to dislodge the receiver's code tracking loop. Instead, the meaconer will be forced to jam the target receiver to force re-acquisition, which will alert the target to the attack. In any case, GNSS system designers have an incentive to make T_W as small as possible to increase the difficulty of a meaconing attack.

Security Code Estimation and Replay Attack

A Security Code Estimation and Replay (SCER) attack allows greater flexibility than a meaconing attack in manipulating the target receiver's PVT solution. In a SCER attack, a spoofer receives and tracks individual authentic signals and attempts to estimate the values of each signal's security code on-the-fly. It then reconstitutes a consistent ensemble of GNSS signals, with the security code estimates taking the place of the authentic security codes, and transmits the ensemble toward the target receiver. For a single GNSS signal corresponding to a particular satellite, the combined SCER-spoofed and authentic received signals can be modeled as (16.1) but with $r_S(t)$ as in (16.12) and

$$r_1(t) = \alpha \sqrt{P_S} \hat{W}(t - \tau_c) X[\tau_c, \theta_c(t)] + n_c(t),$$

where $\hat{W}(t - \tau_c)$ represents the security code estimate arriving with a delay of $d = \tau_c - \tau > 0$ seconds relative to the authentic security code $W(t - \tau)$, $n_c(t)$ is noise introduced by the spoofer (e.g., due to quantization effects in the signal generation), and other quantities are as introduced previously. The delay d can be modeled as the sum $d = p + e$ of a processing and transmission delay $p > 0$ and an estimation and control delay $e > 0$. The delay p represents the combined minimum signal processing delay and additional propagation time and

does not contribute to better estimates of the security code chips. The delay e represents an additional delay imposed by the spoofer to improve its estimate of the security code chip values and to control the relative phasing of the spoofing signals so as to impose spoofer-defined position and timing offsets on the defender.

Mounting a stealthy SCER attack is challenging if the target receiver has been designed to detect SCER spoofing. The attacker must keep $d = p + e$ small enough to remain within the target receiver's clock uncertainty but must extend e enough to reliably estimate the security code chip values. The following two SCER attack strategies serve to illustrate this tradeoff.

Zero-Delay Attack. Consider a spoofer that is co-located with the target GNSS receiver's antenna and has negligible processing delay so that $p \approx 0$. Assume that $e = 0$, meaning that the spoofer adds no estimation and control delay. Thus, $d = p + e \approx 0$. In this zero-delay attack, $\tau_c \approx \tau$, which implies that each spoofing signal is approximately code-phase-aligned with its authentic counterpart as received by the target receiver.

Despite such code phase alignment, a zero-delay attack can still alter the target receiver's position and time by injecting false messages through $D(t)$ (e.g., erroneous satellite ephemeris or clock model parameters or an erroneous leap second). However, with $e = 0$, the spoofer's security code estimate $\hat{W}(t)$ will be highly erratic for the first few microseconds following an unpredictable chip transition in $W(t)$. This is illustrated in Fig. 16.14, which shows simulated time histories of $\hat{W}(t)$ for two different chip value estimation strategies over the first 20 μs after the beginning of a security code chip with $T_W > 20 \mu\text{s}$. In this scenario, for which the spoofer C/N_0 is an unusually high 54 dB Hz, the spoofer's chip estimates become reliable after about 8 μs . For each 3 dB drop in spoofer C/N_0 , the interval required for reliable chip estimates doubles.

The key to zero-delay SCER attack detection, as explained in [16.54], is to develop a detection statistic that is sensitive to the increased error variance in $\hat{W}(t)$ in the crucial early moments immediately following unpredictable transitions in $W(t)$.

Nonzero-Delay Attack. In a nonzero-delay SCER attack, the spoofer rebroadcasts a counterfeit signal that arrives at the defender's RF front end with a delay $d > 0$ relative to the authentic signal. Any significant delay d (e.g., greater than about 20 ns) in the spoofer's counterfeit signal at the beginning of an attack would be immediately obvious to a target receiver that has been continuously tracking authentic signals since before the beginning of the attack. Therefore, the spoofer's strategy in the nonzero-latency SCER attack is typically to break the target receiver's tracking continuity by jamming or blocking the authentic signals for an interval of time before initiating the spoofing attack, thus, widening the target receiver's timing uncertainty, or *window of acceptance* [16.53, 55, 61]. The required duration of the signal-denial interval depends on the desired delay d and on the assumed stability of the target receiver's clock (for stationary receivers) or clock and inertial measurement unit (for moving receivers). For the low-cost temperature-compensated crystal oscillators (TCXOs) typical in commercial GNSS equipment, in-the-field stability is approximately 10^{-7} . Ovenized crystal oscillators (OCXOs), common in more demanding timing applications, have stability of approximately 10^{-10} . Thus, widening a TCXO-driven static target receiver's time uncertainty by 8 μs would require approximately 80 s of jamming or blockage, and widening an OCXO-driven static receiver's time uncertainty by the same amount would require approximately one day of jamming or blockage.

After the jamming-or-blockage prelude, the non-zero-delay SCER attacker initiates a spoofing attack in which d can be as large as the target receiver's tim-

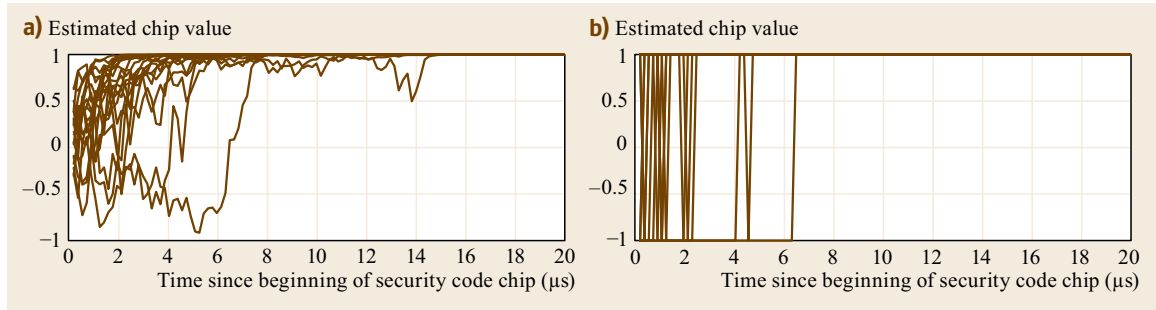


Fig. 16.14a,b Simulated time histories of security code chip estimates $\hat{W}(t)$ for a minimum mean square error (MMSE) estimator (a) and for a maximum a posteriori (MAP) estimator (b) over the first 20 μs after the beginning of a unity-valued security code chip for a spoofer with received $C/N_0 = 54$ dB Hz (after [16.54], courtesy of IEEE)

ing uncertainty. The attacker exploits the component e of this delay to more accurately estimate the value of each unpredictable chip in $W(t)$ so that $\hat{W}(t)$ appears accurate to the target receiver. Long security code chips (e.g., $T_W = 40$ ms as suggested for civil navigation message (CNAV) NMA in [16.54, 56]) allow the spoofer to significantly increase e and thereby generate highly accurate chip estimates. However, a large delay $d = p + e$ is itself a liability for the spoofer because of the long jamming-or-blockage interval required. Thus, the spoofer finds itself vulnerable to detection at low d due to poor security code chip estimates and at high d due to a noticeable timing delay.

Note that, with a SCER attack, the attacker can eventually specify an arbitrary position and an arbitrary delayed time as the spoofer slowly pulls each signal's code phase to the desired offset. Note also that if $W(t) = 1$ (i.e., the GNSS signal is not security enhanced), then the attacker need not delay at all: He can exploit the near-perfect predictability of $D(t)$ to anticipate the next navigation data symbol value and ensure that it arrives at the target receiver's antenna just on time – perfectly aligned with the true $D(t)$ [16.58]. Thus, the unpredictability of the security code – even a low-rate code such as in NMA – forces a SCER spoofer to expose himself with a jamming-or-blockage

attack prelude. Finally, note that signal jamming or blockage for any significant interval of time (relative to the receiver clock stability) must be viewed not only as a temporary nuisance but also as a security threat that persists even after the interference apparently subsides. This is because, in the absence of some other means of verifying the authenticity of GNSS signals, a SCER attack detector's probability of detection is irrecoverably reduced by a loss of signal continuity [16.55].

Effect of Coherence

In a spoofing attack, the complex correlator output modeled in (16.3) contains a desired component $S(t) \equiv r_S^*(t)l(t)$ and an interference component $I(t) \equiv r_I^*(t)l(t)$, both of which are dependent on the local replica's code phase $\hat{\tau}$ and carrier phase $\hat{\theta}$. Denote these as $S(t, \hat{\tau}, \hat{\theta})$ and $I(t, \hat{\tau}, \hat{\theta})$. Also, for a given authentic and spoofing signal pair $r_S(t)$ and $r_I(t)$, let $\varphi(t) \equiv \theta_c(t) - \theta(t)$ be the relative carrier phase.

If a spoofing attack is code-phase aligned so that $|\tau_c - \tau| < T_C$, and Doppler matched so that

$$\frac{1}{2\pi} \left| \frac{d\varphi}{dt} \right| < \frac{1}{T_a}$$

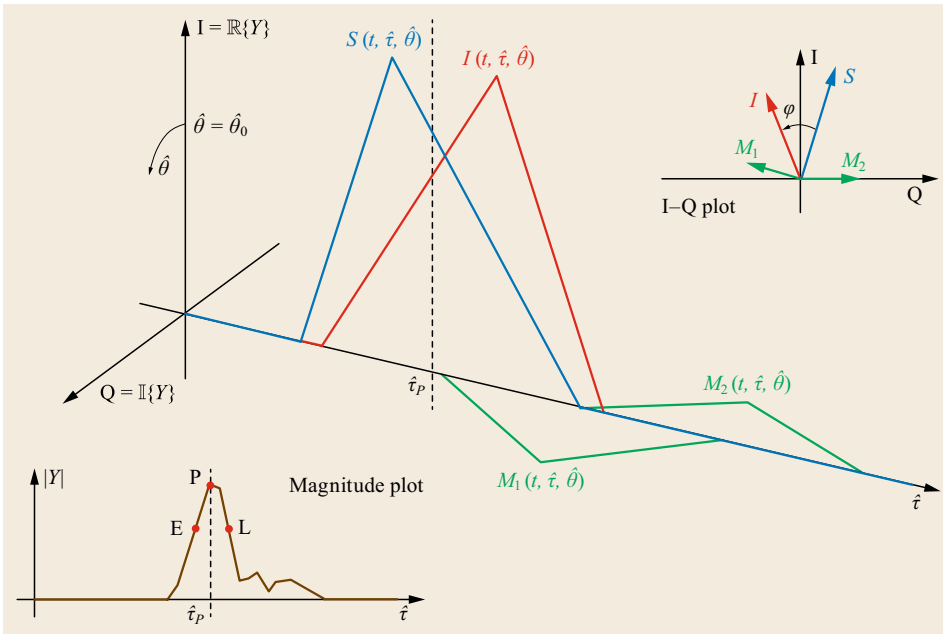


Fig. 16.15 Stylized complex correlation functions depicting a spoofing attack in which $|\tau_c - \tau| < T_C$ and $d\varphi/dt \approx 0$. The blue trace marked $S(t, \hat{\tau}, \hat{\theta})$ represents the desired signal correlation function, the red trace marked $I(t, \hat{\tau}, \hat{\theta})$ represents the interference (spoofing) signal correlation function, and the green traces marked $M_i(t, \hat{\tau}, \hat{\theta})$, $i = \{1, 2\}$, represent two multipath correlation functions. The receiver's code and carrier tracking loops track the composite correlation function, $Y(t, \hat{\tau}, \hat{\theta})$, whose magnitude is shown in the lower inset plot along with the early, prompt, and late correlation taps

with T_a is the accumulation interval from Fig. 16.2, then $r_S(t)$ and $r_I(t)$ are substantially frequency coherent and thus cannot be considered statistically independent. As a consequence, the combined signal power P_T is not simply the sum $P_T = P_S + P_I + P_n$, as in (16.2), but depends on $\tau_c - \tau$, φ , and the relative spoofing amplitude α . Figure 16.15 shows the relationship between $S(t, \hat{\tau}, \hat{\theta})$ and $I(t, \hat{\tau}, \hat{\theta})$ in this regime.

The interference power P_I can be decomposed as $P_I = \alpha^2 P_S + P_{nc}$, where P_{nc} is the power in the noise component $n_c(t)$. If code-phase alignment and Doppler matching are approximately achieved in a spoofing attack ($|\tau_c - \tau| \approx 0$ and $d\varphi/dt \approx 0$), the possibility of which was demonstrated in [16.50] against a nonsecurity-enhanced GNSS signal, then P_T can be expressed as

$$P_T = \left[\sqrt{P_S} + \sqrt{\alpha^2 P_S} \cos(\varphi) \right]^2 + \alpha^2 P_S \sin^2(\varphi) + P_{nc} + P_n. \quad (16.13)$$

This expression indicates that the noise components P_{nc} and P_n , which are noncoherent with the authentic signal, add directly to P_T , as does $\alpha^2 P_S \sin^2(\varphi)$, which is the power in the spoofing signal's frequency-coherent component that lies in phase quadrature to the authentic signal. By contrast, $\alpha^2 P_S \cos^2(\varphi)$, which is the spoofing power component that is phase aligned with the authentic signal, does not add directly to P_T but instead interacts with the authentic signal as shown. For $k \in \mathbb{Z}$, the spoofing signal contributes maximally to P_T when $\varphi = k2\pi$ (phase alignment), minimally when $\varphi = (1+2k)\pi$ (antiphase alignment), and power-additively – as if it were a purely noncoherent signal – when $\varphi = (1/2+k)\pi$ (orthogonal alignment).

It is interesting to note that if φ is treated as a random variable uniformly distributed on $[0, 2\pi]$, then the expected value of P_T is equivalent to the P_T that arises in the case of purely noncoherent interference signals; that is, $E[P_T] = P_S + P_I + P_n$. Hence, for an ensemble of statistically independent spoofer-and-authentic signal pairs, (16.2) remains a useful approximation for the power contributed by each pair even when the spoofer can achieve Doppler frequency alignment ($d\varphi/dt = 0$) but has no finer control over the carrier phase. By distinction, if the spoofer has knowledge of the target

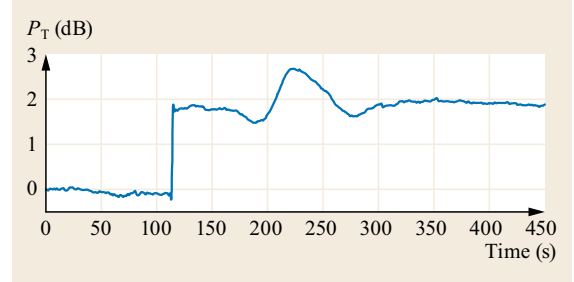


Fig. 16.16 Total received power P_T in a 2 MHz band centered at the GPS L1 frequency showing the onset of a spoofing attack using the testbed described in [16.62], normalized by the average value of P_T prior to the attack. The attack begins with a sudden increase in P_T just before 100 s. Thereafter, the total authentic signal power and total spoofing power were maintained constant; thus, the oscillations in P_T are due to the frequency coherence between the spoofing and authentic signals, with each pair of spoofing-and-authentic signals having similar values of φ

receiver's antenna position to within a small fraction of a carrier wavelength, then it can arbitrarily adjust α and φ to exercise full control over P_T according to (16.13). Figure 16.16 demonstrates that frequency-coherent spoofing signals affect P_T as expected.

An important consequence of a spoofer's having arbitrary control over α and φ is that, by choosing $\alpha = 1$ and $\varphi = \pi$ for each spoofing and authentic signal pair, a spoofer can effectively annihilate the authentic signals at the location of the target antenna. Such a nulling attack has the effect of jamming the target receiver while *reducing* the total received power P_T in the GNSS band of interest. Moreover, the nulling signals could be paired with an independent ensemble of spoofing signals to simultaneously eliminate the authentic signals while presenting clean counterfeit signals to the target receiver. The attacker could thus evade tests, such as the received power test proposed in [16.35] and the pincer defense proposed in [16.63], designed to detect anomalies in the total received power or distortion in the correlation function caused by interaction of the authentic and spoofing signals. GNSS antennas that are clearly visible to the public from close range and those whose coordinates are publicly posted to subdecimeter accuracy are at greatest risk of such nulling attacks.

16.6 Interference Detection

Many schemes for detecting and mitigating GNSS interference have been proposed since the early days of GPS. These schemes apply at one or more of three application points in the GNSS signal processing chain, as shown in Fig. 16.17: (1) the analog stage, (2) the post-digitization but precorrelation stage, and (3) the correlation and post-correlation stage. Several effective interference detection schemes are detailed in this section; the following section treats interference mitigation.

16.6.1 C/N_0 Monitoring

A drop in a receiver's measured C/N_0 on any channel that cannot be explained by signal shadowing indicates interference of some type. C/N_0 is related to the SNR of the complex accumulations Y_k (Fig. 16.2) on which code and carrier tracking are based by $\text{SNR} = CT_a/N_0$. As C/N_0 measurements are generated post-correlation, C/N_0 monitoring applies at point (3) in Fig. 16.17.

Given measured C/N_0 , one can be assured that code and carrier tracking will perform no better than what would be expected for $\text{SNR} = CT_a/N_0$. Nominal C/N_0 values across all tracking channels do not, however, guarantee the absence of interference, since spoofing interference, whether intentional or not, can cause the affected receiver to report perfectly normal C/N_0 values. For example, the spoofer described in [16.62] can dictate the received C/N_0 for each signal by adjusting the relative magnitudes of its output signals and adding artificial noise to the signal ensemble.

Given that C/N_0 loss is often caused by signal shadowing, and that nominal C/N_0 values are no guarantee of the absence of interference, a C/N_0 monitor such as proposed in [16.64] is best applied in combination with other complementary techniques for GNSS interference.

16.6.2 Received Power Monitoring

Monitoring the total received power P_T in a GNSS band of interest, known as received power monitoring

(RPM), is one of the simplest and most effective strategies for detecting interference [16.35, 65, 66]. For systems with multibit-quantized sampling and automatic gain control (AGC) in the RF front end, estimating P_T is as easy as measuring the voltage applied by the AGC unit to adjust the signal amplitude before quantization. In a constant-gain system with sufficient dynamic range to prevent quantization saturation, P_T can be estimated directly from the precorrelation samples. In any case, RPM can be thought of as applying at point (2) in Fig. 16.17.

Figure 16.18 shows the nominal power spectrum about the GPS L1 frequency as measured at the output of a high-quality GNSS antenna and front-end system. Despite their statistical independence and low power, the received GPS L1 C/A signals combine to yield an obvious enhanced density in the familiar $\text{sinc}^2(fT_C)$ pattern near L1 that rises above the noise floor.

For interference detection with a suitably low false alarm rate, one must examine the size and predictability of variations in P_T that can be considered natural or otherwise innocuous. Figure 16.19 shows a two-day record of P_T for the setup in Fig. 16.18 in the 2 MHz band centered at L1. The time history reveals marked diurnal variations, the result of diurnal patterns in temperature, solar radiation, and the overhead satellite constellation. Even though the record's diurnal repeatability is evidently only good to approximately 0.3 dB, its predictability given knowledge of local temperature and satellite orbital ephemerides is actually better than this.

Figure 16.20 offers an expanded view of a 7.5 min interval using the same setup and showing both the 2 and 10 MHz traces. The different size of the variations in the two traces at time scales less than about 150 s indicates that the variations do not originate in broadband noise; they are likely due to multipath effects at the carrier-phase level caused by reflections off nearby surfaces and by atmospheric diffraction and refraction. Close examination of multi-day records of P_T reveals that these short-time-scale variations do not repeat appreciably at the solar or sidereal day. In sum-

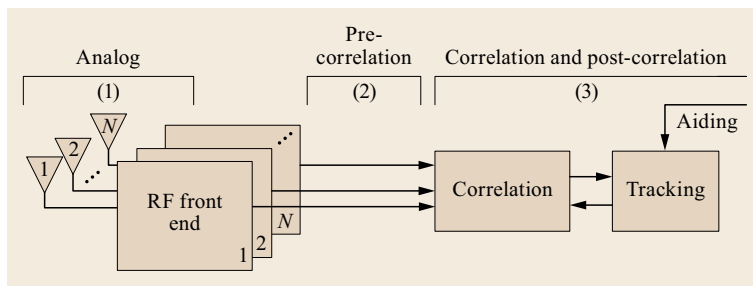


Fig. 16.17 Application points for interference detection and mitigation: (1) in the analog stage prior to digitization, (2) after digitization but before correlation, and (3) in correlation and in post-correlation tracking and PVT estimation

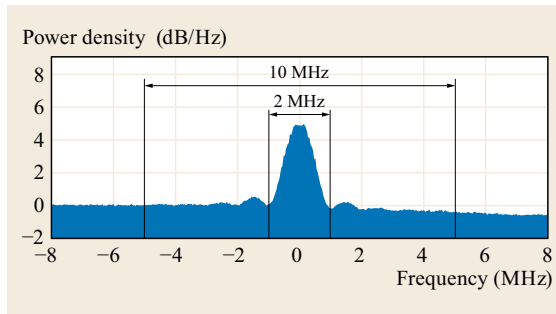


Fig. 16.18 Power spectrum centered at the GPS L1 frequency as estimated from a 1 s interval of data captured via a high-quality static antenna and RF front-end combination in a moderately quiet outdoor RF environment. Bands for 2 and 10 MHz power measurements are shown. The power density scale has been centered near the noise floor for ease of viewing. In absolute units, the noise floor sits at approximately -204 dBW/Hz

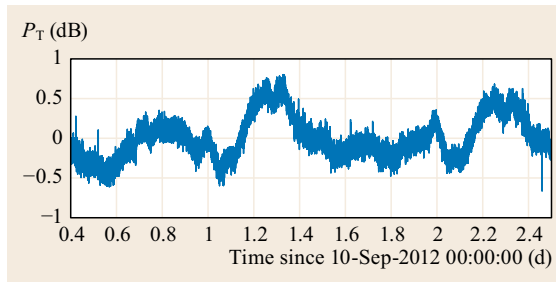


Fig. 16.19 A two-day record of received power P_T in the 2 MHz band shown in Fig. 16.18, normalized by the average received value over the interval

mary, it appears that for a static antenna, the practically unpredictable variations in P_T about L1 have root-mean-squared deviations of at least 0.1 dB for a 2 MHz band and 0.05 dB for a 10 MHz band.

For a dynamic antenna, P_T can be much more variable. Figure 16.21 shows a time history of P_T for a receiver mounted on a vehicle driving through the streets of downtown Austin, Texas. The P_T excursions, the largest of which exceeds 1 dB, would be unpredictable to a GNSS user without an up-to-date RF interference map of the area.

Against background variations that are unpredictable at the 0.1 dB level, or even the 1 dB level, deliberate jamming from close range remains obvious, as revealed by the effect on P_T of highway motorists using PPDs shown in Fig. 16.22. Naive spoofing also has an obvious effect: consider the sudden 2 dB uptick of P_T in Fig. 16.16. However, contrary to the claims in [16.35], RPM is not a generally effective means of detecting spoofing. This is because the increase in P_T

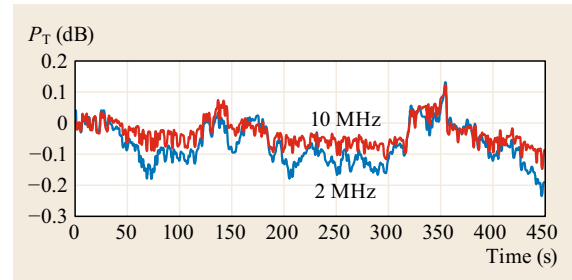


Fig. 16.20 A 7.5 min record of received power in the 2 and 10 MHz bands shown in Fig. 16.18, normalized by the initial values of P_T in each band

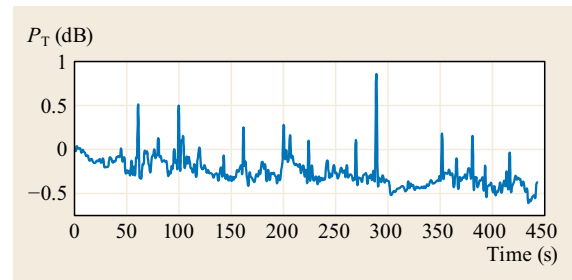


Fig. 16.21 Received power P_T in a 2 MHz band centered at the GPS L1 frequency averaged over 1 s intervals for a receiver mounted on a vehicle driving through the streets of downtown Austin, Texas. The data correspond to the *clean dynamic* data record from [16.67]

during a spoofing attack may be smaller, or not significantly larger, than unpredictable variations in P_T due to causes other than spoofing. As mentioned in Sect. 16.5.2, a spoofer able to arbitrarily control the relative amplitude α and phase ϕ of each spoofing signal can annihilate the authentic signals and supplant them with counterfeit signals of equal power, thereby, maintaining P_T constant.

A spoofer lacking precise control over ϕ cannot prevent an increase in P_T while successfully capturing the target receiver's tracking loops, but the increase in P_T can be small: For a commercial-grade GNSS receiver, the uptick in P_T may be as small as 0.56 dB [16.62]. If unpredictable natural variations in P_T are modeled as a Gaussian process with a 0.1 dB standard deviation and a 150 s decorrelation time, then a detection threshold equal to $\gamma = 0.44$ dB would be sufficient to detect such an uptick with high probability while maintaining a once-per-year false alarm rate. However, the natural variations in P_T have a much thicker high-side probability distribution tail than a Gaussian process. For example, as detailed in Table 16.2, solar radio bursts would cause P_T to exceed $\gamma = 0.44$ dB every 9.2 days on average during solar maximum. Note that

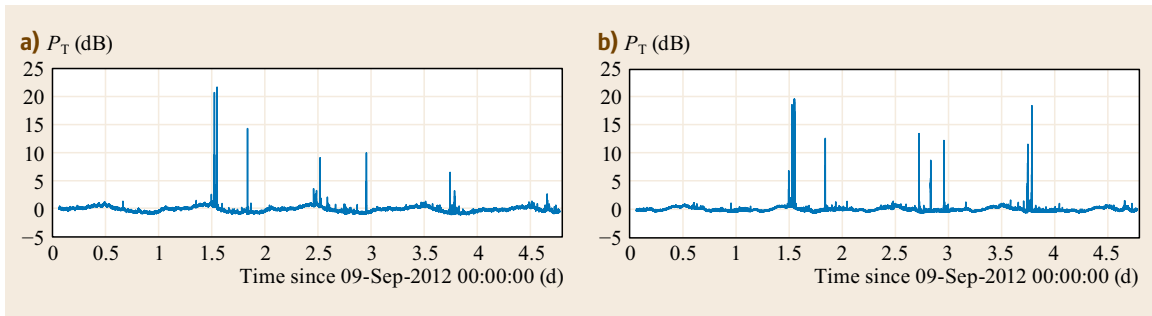


Fig. 16.22a,b Received power in the 10 MHz band centered at GPS L1 at two sites 1 km apart that straddle State Highway 1, west of Austin, TX. **(a)** Data from site located at the Center for Space Research. **(b)** Data from site located at Applied Research Laboratories. Both traces are normalized by the average value of P_T over the interval. The large excursions in P_T are due to motorists using PPDs as they travel along the highway

although spoofing alarms could be dismissed during known solar radio burst events, which can be independently monitored – even predicted [16.68], this offers little protection, for a clever attacker could time his attack to coincide with the arrival of a sizable burst.

Besides solar radio bursts, nonspoofing interference endemic in urban environments and near major thoroughfares can often cause an increase in P_T exceeding $\gamma = 0.44$ dB, as shown in Figs. 16.21 and 16.22. One might argue that it is perfectly appropriate for a spoofing detector to alarm in the presence of a solar radio burst or an intentional jammer, but the consequences of spoofing can be much more malign than those of natural interference or jamming, and so it behooves a defender to distinguish between these.

16.6.3 Augmented Received Power Monitoring

When acting alone, RPM is effective at detecting strong interference but cannot be considered a reliable detector of weak interference such as low-power spoofing. It can, however, be paired with other tests that are sensitive to GNSS-like structure in the received signal to yield a powerful joint detection test for spoofing, provided the spoofer cannot arbitrarily manipulate α and φ . Three RPM augmentation strategies are discussed in the following sections.

Augmentation with C/N_0 Monitoring

A simple C/N_0 monitor will not detect spoofing signals whose C/N_0 values are matched to those of the authentic signals. But when paired with RPM, C/N_0 monitoring becomes a reasonably reliable detection strategy because it is challenging for a spoofer to ensure nominal received C/N_0 values without significantly increasing P_T . Only with a nulling attack, such as described in

Sect. 16.5.2, can a spoofer ensure that C/N_0 matching does not increase P_T . Without nulling, C/N_0 matching (with no unusual variations) requires overwhelming spoofing power, which manifests as increased P_T .

Augmentation with Precorrelation Structural Power Content Analysis

The precorrelation structural power content analysis method advanced in [16.69] detects the presence of spoofing based on the excessive power content of GNSS-like signals in the received raw samples. In the absence of RPM, a spoofer can evade this detector by transmitting with overwhelming power, thus, driving the received authentic signals into the noise floor as the receiver's AGC compensates for the high received total power. The method of [16.69] will then only measure precorrelation structural power content commensurate with a single signal for each expected received waveform, and will thus fail to alarm. However, when combined with RPM, a structural power detector becomes powerful for spoofing detection. As for C/N_0 monitoring, augmentation with RPM forces the spoofer to either mount a nulling attack or be exposed with high likelihood in the joint test statistic.

Augmentation with Distortion Monitoring

The pincer defense advanced in [16.63] thoroughly embraces the concept of augmenting RPM for improved spoofing detection. Its name is meant to evoke a pin-cering, or trapping, of the spoofing signals between an RPM and a signal distortion monitor. As with C/N_0 and precorrelation structural power monitoring, distortion monitoring acting on its own cannot detect a spoofing attack executed with overwhelming power because the interaction between the authentic and false signals, which is the source of the signal distortion sought, is eliminated by action of the AGC as the spoofing-to-authentic power ratio increases.

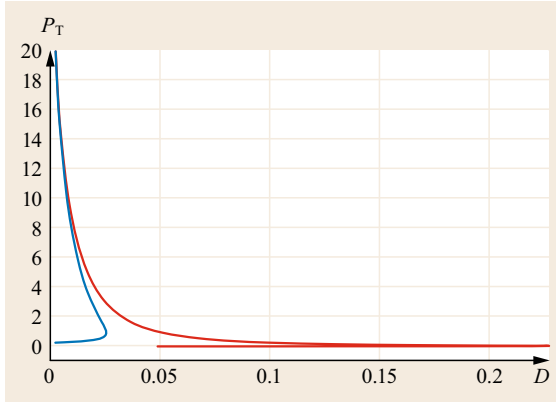


Fig. 16.23 Distortion (in the same units as accumulation), as a function of P_T for in-phase (blue) and antiphase (red) multipath or spoofing interference at a fixed delay of 0.15 chips. For the same delay, all other relative phases yield distortion profiles that lie within this envelope (after [16.63]; reprinted with permission)

The GNSS signal quality monitoring literature has proposed several metrics for signal distortion [16.70]. These metrics are all calculated based on correlation products and so apply at point (3) in Fig. 16.17. The pincer defense adopts the so-called symmetric difference D . Let Y_E and Y_L be the early and late complex accumulations with a predetermined early late spacing, respectively. Then, D is defined as the magnitude of the complex early-late difference: $D \equiv |Y_E - Y_L|$. Thus, D is sensitive to early-late asymmetry in both magnitude and phase.

Unless a spoofer is capable of a nulling attack, then distortion caused by the interaction between authentic and spoofing signals of comparable amplitude will be evident as $D > 0$. Figure 16.23 shows that D approaches zero in the limit of both weak and powerful spoofing. But weak spoofing affects a GNSS receiver no more than multipath, and powerful spoofing can be detected by a significant increase in P_T . Such is the basic premise of the pincer defense.

The pincer defense seeks to classify interference as either spoofing, jamming, or multipath, and to distinguish these categories from normal thermal noise, all on the basis of D and P_T . The challenge can be appreciated in reference to Fig. 16.24, which shows a scatter plot of D and P_T values under simulated spoofing (red), jamming (blue), multipath (black), and clean (only thermal noise; green). Clearly, there is overlap between the categories, especially between low-power spoofing and severe multipath.

The pincer defense detection and identification problem can be stated as follows. Given a time history of measurements $\mathbf{z}_k \equiv [D_k, P_{T,k}]^\top$, $k \in \mathcal{K} \equiv \{1, 2, \dots, N\}$,

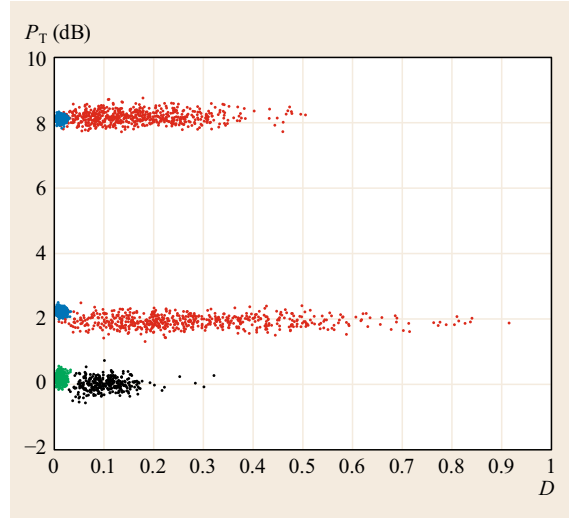


Fig. 16.24 Scatter plot showing simulated D and P_T for clean (only thermal noise; green), multipath (black), spoofing (red), and jamming (blue) scenarios. The spoofing and jamming scenarios are simulated at two different power levels. The simulated accumulation amplitudes were chosen so that D was allowed to range from 0 to 1 (after [16.63]; reprinted with permission)

determine whether the receiver experienced no interference (the null hypothesis, H_0), or, whether for $k \in \mathcal{K}_I \equiv \{k \in \mathcal{K} | k \geq k_o\}$, the receiver experienced multipath (H_1), jamming (H_2), or spoofing (H_3), where k_o is the interference onset index. The problem reduces to a set of generalized likelihood ratio tests conditioned on estimates of k_o , on the interference amplitude α , and, for H_2 and H_3 , on an estimate of the code delay τ_c .

Figure 16.25 shows an example observation space for a single measurement \mathbf{z}_k , partitioned into decision regions for the four hypotheses. The region boundaries depend on the estimates of α and τ_c , on the cost of deciding H_i when H_j is true, $i, j \in \{0, 1, 2, 3\}$, and on the prior probabilities of the four hypotheses.

The problem formulation introduced above is not unique to the pincer defense; indeed, the detection and identification problem for all interference detection techniques can be formulated in terms of H_0 , H_1 , H_2 , and H_3 . Joint detection and classification offer the dual benefit of increased detection power and actionable information about the nature of the interference; these benefits, however, come at the cost of additional computational complexity [16.71].

16.6.4 Spectral Analysis

If the discrete-time quantized samples produced by a receiver's RF front end are accessible to a module capable

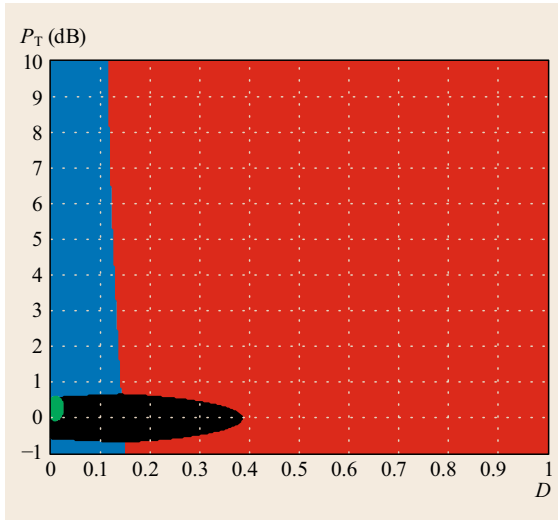


Fig. 16.25 Example observation space for a single measurement \mathbf{z}_k divided into decision regions for clean (only thermal noise; green), multipath (black), spoofing (red), and jamming (blue) (after [16.63]; reprinted with permission)

of performing a discrete Fourier transform (DFT), then the received signal power spectrum can be periodically estimated and analyzed. On multifrequency receivers, this may entail analysis of six or more individual GNSS bands. The computational burden of such analysis can be reduced by use of an efficient DFT implementation and by extending the interval between production of power spectra.

Power spectrum analysis is both a simple and powerful interference diagnostic technique, indicating not only the presence but also the nature of interference, whether wideband or narrowband, constant or fleeting. Figure 16.18 shows the power spectrum centered at L1 produced by a 1 s interval of data from a high-quality static receiver in a quiet RF environment. The spectrum shown is an estimate based on the usual periodogram technique of averaging the spectra produced by overlapping sections of the original data, with each time segment weighted by a windowing function.

The key challenge of interference detection and identification via power spectral analysis is distinguishing actual interference from spectral variability due to signal shadowing, multipath, temperature variation, and the changing overhead GNSS signal constellation. As shown in the example data set in Fig. 16.19, the aggregate power in the 2 MHz band centered at L1 can vary by more than 1 dB even when no interference is present. Much of this variation is periodic and therefore predictable. Sophisticated spectral analysis techniques could apply models or machine learning to distinguish

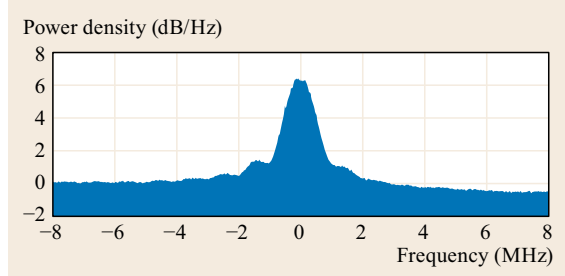


Fig. 16.26 Power spectrum under the same conditions as Fig. 16.18 except that the receiver is now subject to a GPS spoofing attack using the testbed described in [16.62]

novel interference from background variability. Naturally, the problem is much less challenging for static receivers than for mobile ones.

Spectral analysis, even acting alone, can be effective at discovering spoofing. Figure 16.26 shows the same 16 MHz wide power spectrum as in Fig. 16.18 and for the same receiver but for data captured during a spoofing attack in which a false signal was generated for each authentic signal. The profile in Fig. 16.26 thus represents the power spectrum of an admixture of spoofing and authentic signal ensembles. The attack was designed to be stealthy, achieving approximate authentic signal nulling (as described in Sect. 16.5.2) during the interval of data from which the spectrum was computed. Even so, obvious differences are evident between Figs. 16.26 and 16.18. Besides the approximately 2 dB increase in power in the 2 MHz band centered at L1, the side lobes on both sides of the main lobe are more prominent in the spoofed spectrum. Such differences offer hope that a useful degree of spoofing detection could be provided based solely on power spectral measurements.

16.6.5 Cryptographic Spoofing Detection

A GNSS signal modulated with an unpredictable but verifiable security code $W(t)$, as in (16.12), is much more resistant to spoofing than a GNSS signal with no purposeful unpredictability. The security code $W(t)$ is best implemented as a cryptographic sequence. In NMA, $W(t)$ is a low-rate (e.g., 50–250 Hz) binary sequence containing periodic digital signatures that are unpredictable at transmission but can be verified upon receipt to certify the origin of the complete data sequence $D(t)$ [16.55–57]. Alternatively, $W(t)$ can be implemented as a high-rate (e.g., 500–10 000 kHz) binary sequence whose chip interval can be as short as that of the underlying spreading code $C(t)$, as is the case for the GPS Y and M signals, the Galileo PRS signal, and spread-spectrum security codes proposed for civil applications [16.53].

The security of the military GPS Y and M codes is based on symmetric-key cryptography. The GPS control segment generates a pseudorandom binary spreading code sequence based on a combination of secret keys. A military receiver generates a local replica of the same sequence based on a functionally equivalent set of secret keys, enabling despreading and signal tracking. Unauthorized agents are presumably denied access to the secret keys, so, in theory, they can neither generate nor predict the spreading sequence, which means they can neither track nor anticipate the military GPS signals for purposes of spoofing.

It is neither practical nor prudent to base civil security codes on symmetric-key cryptography. Instead, all proposed civil schemes are based on public-key cryptography or on delayed disclosure of secret keys. Even the technique proposed in [16.72], which leverages the military Y code to secure civil GPS receivers, assumes that the Y code is revealed to the receiver some time after receipt.

Detection

Spoofing of a security-code-enhanced GNSS signal is easily detected if the counterfeit signal's security code fails digital signature verification (for low-rate security codes) or fails to generate significant power when correlated against a replica security code (for high-rate security codes). Only meaconing and SCER attacks are capable of generating counterfeit signals that could satisfy these preliminary tests.

For both meaconing and SCER attacks, the detection techniques discussed previously can be quite effective, particularly augmented received power monitoring and spectral analysis. For SCER attacks, another powerful tailored detection test can be formulated [16.54, 55]. The test's decision statistic is based on received power P_T and on a specialized correlation statistic L . Given its dependence on P_T , SCER attack detection can be thought of as another type of received power monitoring augmentation, much like C/N_0 monitoring or the pincer defense.

The SCER attack detector's specialized correlation statistic L is designed to be sensitive to the high error variance of the spoofer's security code estimate $\hat{W}(t)$ in the moments immediately following each unpredictable chip transition. Reference [16.54] develops the statistic and describes its distribution under H_0 (no attack) and H_1 (SCER attack). What follows briefly describes how the statistic is generated within a receiver and offers an example test result.

Let W_k be the value of the security code $W(t)$ during the k -th chip. For convenience, assume that the receiver's accumulation interval is equivalent to the length of W_k , as for NMA. Then, the correlation statistic L can be generated as shown in Fig. 16.27. The lower signal path is the standard matched-filter-type correlation operation previously depicted in continuous time in Fig. 16.2. The product of the incoming samples r_i and a complex local signal replica $l_i = W_k C_l(t_i - \hat{t}_i) \exp(-j(2\pi f_{IF} t_i + \hat{\theta}(t_i)))$ is accumulated over the interval spanned by W_k to produce the prompt complex correlation products $I_k + jQ_k$ that get fed to code and carrier tracking loops. The code tracking loop also ingests correlation products from identical paths – not shown – involving early and late versions of $C_l(t_i - \hat{t}_i)$.

The upper path in Fig. 16.27 produces the SCER attack detection statistic L . The real part of the product $r_i l_i$ is multiplied by a smooth weighting function $\beta(n_{ki})$, defined in [16.54], that gives full weight to the i_k -th sample but decays rapidly toward zero for subsequent samples. This weighting has the effect of suppressing those samples over which the error variance in the spoofer's security code chip estimate \hat{W}_k has become small because the spoofer has had sufficient time to obtain an accurate estimate of W_k ; as illustrated in Fig. 16.14, only the early high-variance samples are useful in distinguishing H_1 from H_0 . The weighted product $\beta(n_{ki}) \mathcal{R}(r_i l_i)$ is accumulated over the interval spanned by W_k to produce the single-chip detection statistic S_k , N of which are biased, squared, and accumulated as shown to produce the final statistic L . The constants a and b are related to the theoretical mean

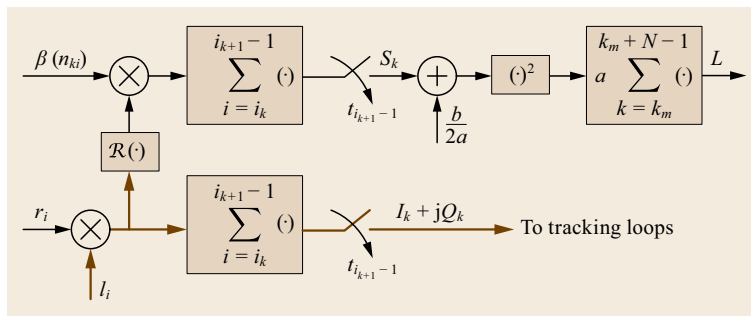


Fig. 16.27 Block diagram illustrating how generation of the SCER attack detection statistic L relates to standard GNSS signal correlation. *Thick brown lines* denote complex signals, whereas *thin black lines* denote real-valued signals

μ_p and variance σ_p^2 of S_k under H_p , $p = 0, 1$ by

$$a = \frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}, \quad b = 2 \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2} \right).$$

SCER Attack Detection Example

The test results shown in Fig. 16.28 are expressed in terms of the empirical distribution of L at various stages of an example SCER attack performed in the testbed of [16.62]. The top panel shows the attack prelude during which only the authentic signal is present. At this stage, the histogram of L values exhibits good correspondence with the theoretical null-hypothesis probability distribution $p_{L|H_0}(\xi|H_0)$, where ξ is the value at which the probability density of the detection statistic L is evaluated. The center panel shows the situation during the initial stage of the attack when the authentic and spoofing signals are aligned to within a small fraction of the $\approx 1 \mu\text{s}$ spreading code chip interval. Because the counterfeit and authentic signals in this test are so nearly matched in power, this stage manifests strong interaction between the two in the defender's complex-valued prompt correlator. Such interaction violates the either/or assumption of the SCER detection test. The detection statistic does exceed the threshold more than half the time, but instead of clustering within $p_{L|H_1}(\xi|H_1)$, it exhibits spreading driven by variations in the relative carrier phase of the interacting authentic and spoofing signals.

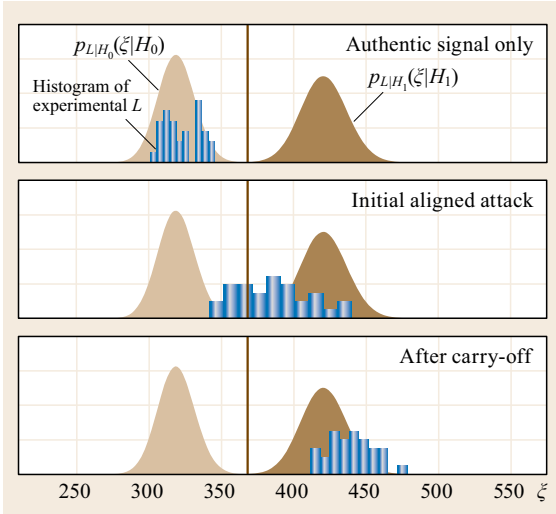


Fig. 16.28 Histograms of experimentally generated detection statistics L (bar plots) compared with the detection threshold (thick vertical line) and the theoretical distributions $p_{L|H_j}(\xi|H_j)$, $j = 0, 1$ at various stages of a zero-delay SCER attack

After the spoofer has successfully carried off the defender's tracking points and the authentic and spoofed correlation peaks are separated by more than two spreading code chips, the SCER detector's attack model again becomes valid. The bottom panel of Fig. 16.28 shows that at this stage, the detection statistic clearly clusters beyond the detection threshold and roughly within the theoretical $p_{L|H_1}(\xi|H_1)$ distribution.

16.6.6 Antenna-Based Techniques

A GNSS receiver employing only a single, static antenna cannot measure the arrival direction of incoming signals, but a receiver with a moving antenna or multiple antenna elements can discern arrival direction and can use this information to detect interference. Antenna-based techniques are powerful for interference detection because an interference source commonly transmits from a single antenna whereas GNSS signals come from a spatially diverse set of overhead satellites. A spoofing detector based on a single moving antenna is developed in [16.73], and one based on a pair of static antennas is developed in [16.74]. The latter demonstrates nearly immediate spoofing detection with a low-cost system in a live spoofing attack.

16.6.7 Innovations-Based Techniques

A final opportunity for detecting spoofing interference arises in the PVT estimation algorithm that draws in the GNSS pseudorange and carrier-phase observables produced by the tracking loops, or, in the case of a vector tracking architecture, in the consolidated tracking and PVT estimation algorithm. The tracking block in Fig. 16.17 is intended as a generic reference to such tracking and estimation functions, and would be the application point for innovations-based spoofing detection techniques.

PVT estimation algorithms typically employ a model of the receiver dynamics – including clock dynamics – and may have access to non-GNSS aiding data such as from an inertial measurement unit (IMU), barometer, magnetometer, etc. Sequential estimators such as the Kalman filter are commonly used for this purpose, processing a regular cadence of observables and generating a regular output of PVT estimates.

Significant inconsistency between the estimator's predictions and GNSS observables can be detected by standard hypothesis testing applied to the estimator residuals, or innovations (Chap. 24). Reference [16.51] offers a framework for innovations analysis that is optimized for sensor deception, including GNSS spoofing. The framework applies an integrity risk performance

index to account for the fact that a sensor attack only causes harm when the target system exceeds its alert limit – when a ship leaves its assumed transit corridor or a timing system exceeds its required timing accuracy specification, for example. The framework adopts a minimax detection strategy for robustness to unknown spoofer actions. It is shown that an attacker can cause the target system to exceed its protection limits without detection whenever the attack-induced dynamics lie comfortably within the drift envelope of the PVT estimator’s model-based propagation process. For example, PVT estimation based on pseudorange and Doppler observables and inertial sensors, a common combination, can be led astray by a spoofer whose induced error trajectory gradually departs from the true trajec-

tory as if driven by the drift processes in the inertial sensors [16.50].

In response to this vulnerability, [16.75] proposes a powerful detection test for GNSS-guided vehicles that exploits high-frequency platform dynamics caused by environmental disturbances (e.g., wind gusts buffeting an aircraft). These dynamics are practically unpredictable to a would-be spoofer yet easily measured by both the inertial sensors and high-rate (e.g., 20 Hz) carrier-phase observables. An innovations test on the GNSS carrier-phase measurements that exploits such natural dithering, or even purposeful dithering if natural disturbances offer inadequate excitation, poses great difficulty for a spoofer unless the spoofer is physically attached to the target platform.

16.7 Interference Mitigation

GNSS interference detection is the key to avoiding hazardously misleading information in a GNSS-based PVT solution: Once interference has been detected, the user or larger system can make decisions with full knowledge that the trustworthiness of the PVT solution may be compromised. But mere detection does not ensure *continuity* of reliable PVT information, which is a requirement for many systems and users. PVT continuity may be achieved by human intervention: A ship’s crew can fall back to visual, radar, or even celestial navigation once alerted to GNSS interference. But, increasingly, navigation and timing systems are expected to maintain PVT continuity *automatically* in the face of GNSS interference.

One design philosophy gaining traction in recent years views GNSS as so vulnerable to interference that it must be backstopped with an entirely GNSS-independent PVT source. According to this philosophy, the sensible response to detection of threatening GNSS interference is to abandon GNSS, at least temporarily, by failing over to a non-GNSS backup PVT system. But despite impressive advances in IMU and clock stability, in the use of non-GNSS signals of opportunity for PVT, in non-GNSS time distribution, in electro-optical navigation, and in dedicated terrestrial PVT systems, this approach has only proven useful for short intervals of time (a few minutes) or restricted areas of operation (a radius of a few tens of kilometers). So far, GNSS remains irreplaceable because no combination of non-GNSS PVT systems has yet to rival the essential suite of GNSS benefits: (1) global coverage, (2) high PVT accuracy over indefinitely long time intervals, and (3) low cost to users. Accordingly, this section focuses on GNSS interference mitigation techniques that ensure

PVT resilience not by abandoning GNSS but by toughening and augmenting it.

16.7.1 Spectrally or Temporally Sparse Interference

Effective techniques exist for mitigating interference that is sparse in frequency (narrowband) or time (pulsed). Mitigation of spatially sparse interference, that is, interference with a small number of narrow directions of arrival, will be treated in Sect. 16.7.3.

Sparse interference mitigation techniques exploit time correlation in an interference signal’s phase or amplitude to estimate and excise the interference signal, thereby, increasing the desired signal power to noise ratio. The more highly time correlated an interference signal’s amplitude or phase, the more accurately it can be reconstructed and excised, sparing the downstream acquisition and tracking routines from harmful interference effects.

Filtering

Without proper early stage RF filtering, even interference far from GNSS frequency bands of interest can be problematic for a GNSS receiver when the interference is sufficiently strong: The out-of-band signal rejection of the receiving antenna and the first-stage LNA may not be sufficient to prevent a strong out-of-band signal from saturating the LNA. Thus, in mobile handsets and at cellular base stations, one finds GNSS receivers with stringent RF filtering before first-stage amplification despite the direct C/N_0 reduction (equivalent to the filter impedance loss) that such filtering entails.

Narrowband interference within the GNSS band is more challenging to mitigate than out-of-band interference. Selective (high quality factor) analog filtering within a GNSS band of interest requires large and expensive analog filters. Likewise, LNAs with a linear range wide enough to prevent saturation in the face of strong interference are expensive, as are antenna arrays capable of pointing a null toward the interference source. Thus, attenuation of the received signal before low-noise amplification may in some cases be the only economical recourse to prevent LNA saturation. Unfortunately, one pays the full measure of such attenuation in reduced C/N_0 .

Assuming LNA saturation is avoided, properly configured multibit quantization can be a first defense against narrowband interference. As mentioned in Sect. 16.3.2, multibit quantization can yield a conversion gain (an increase in C/N_0 relative to the unquantized discrete-time samples) when the amplitude of the incoming interference is approximately constant. However, for the one-bit (two-level) quantization employed in many low-cost GNSS receivers, quantizer SNR is severely and irrecoverably degraded by the presence of strong narrowband interference. Even two-bit (four-level) quantization may be insufficient to prevent capture of the quantization process by a strong narrowband interferer, if the interference amplitude varies rapidly or if there are multiple narrowband interferers present.

Assuming sufficient quantization resolution, adaptive digital filtering in the precorrelation stage (point (2) in Fig. 16.17) is a low-cost and highly effective way to mitigate in-band narrowband interference. This technique, commonly referred to as adaptive notch filtering, exploits the time correlation of narrowband interference signals to distinguish them from thermal noise and from the desired spread-spectrum signal, both of which look uncorrelated at chip-length sampling intervals.

Adaptive notch filtering can be implemented either as a transversal filter in the time domain or as shaping in the frequency domain. In the time-domain approach, the weights of a transversal filter are adjusted to minimize the filter's output power [16.76]. Solution of the optimal tap weight vector has complexity $\mathcal{O}(n^2)$, where n is the number of samples in the block used to determine the optimal weights. One may trade off performance for reduced computational demand by extending the interval between subsequent computation of the optimal weight vector. Straightforward implementation can yield highly effective interference suppression even for multiple narrowband interferers: *Dimos et al.* [16.77] show that three pure tone interference sources with a combined interference-to-thermal-noise power of 30 dB in the

GPS L1 C/A band can be suppressed by 28 dB. For the same interference power and number of interferers, but with bandwidths of 25, 50, and 100 kHz, suppression performance reduces to 24.25, 20.75, and 16 dB, respectively, showing that time-domain notch filtering performance degrades as the interference bandwidth increases.

The frequency domain approach entails Fourier transformation of a block of n precorrelation samples (possibly weighted by a windowing function), multiplication of the transform by some appropriate filter, and inverse Fourier transformation of the product. The interference suppression filter applied in the transform domain can be generated automatically to whiten the transformed samples. In the simplest approach, regions containing interference peaks exceeding a predefined threshold can be simply blanked out. The transform approach has complexity $\mathcal{O}(n \log(n))$ and so is less computationally burdensome than time-domain notch filtering with continuous updating of the filter tap weighting. Another benefit of the transform approach is that successive transforms can be averaged to produce a power spectrum estimate, which, as mentioned earlier, is a useful tool for general situational awareness of the interference environment.

The distinctive swept tone interference of PPDs can also be considered sparse given its high regularity [16.47]. A model-based technique is developed in [16.49] that effectively estimates the frequency sweep parameters of PPD signals, allowing the interference to be excised. Such model-based filtering is the logical extension of notch filtering for interference signals that are highly predictable and easily distinguished from the desired GNSS signals.

Blanking

Interference signals that are sparse in time, for example, pulsed interference, can be substantially suppressed by so-called pulse blanking [16.1]. Blanking degrades C/N_0 in proportion to the fraction of RF front-end samples that are discarded. A combined adaptive notch filtering and blanking technique is explored in [16.1] to mitigate DME/TACAN interference, which is sparse in both time and frequency.

16.7.2 Spectrally and Temporally Dense Interference

Interference that is both wideband and continuous is spectrally and temporally dense, unlike narrowband or pulsed interference. It may yet be spatially sparse, but a GNSS receiver with a single, static antenna is unable to exploit such sparseness for mitigation. In this section, dense interference will refer to interference

which is both spectrally and temporally dense regardless of its spatial characteristics. The focus will be on signal-processing-based interference mitigation techniques that do not rely on multiple or moving antennas. The next section treats mitigation of spatially sparse interference using multiple or moving antennas.

Dense interference has substantially time-uncorrelated amplitude and phase at the RF front-end sampling rate, making it appear as thermal noise or as a spread-spectrum GNSS signal to the receiver. Spoofing interference (including meaconing) is an example of interference that is especially difficult to mitigate, because by construction it is intended to masquerade as a legitimate GNSS signal. Faced with multiple identically shaped and sized autocorrelation peaks for the same pseudorandom number code, a receiver can easily recognize that a spoofing attack is underway but cannot mitigate the attack – that is, cannot identify and track only the authentic signal – unless the receiver's combined timing and positioning uncertainty is well within the inter-peak separation. For this reason, post-detection mitigation of a subtle spoofing attack is often only possible by exploiting multiple or moving antennas and will therefore be left to the next section.

It is convenient to treat dense nonspoofing interference such as continuous wideband Gaussian interference as if it were thermal noise for purposes of mitigation. Thus, the dense interference mitigation problem becomes identical to the problem of acquiring and tracking weak GNSS signals in an indoor environment except that the multipath effects in the indoor environment are likely to be more severe than in an outdoor interference environment. Mitigation is applied at the correlation and post-correlation stage, or point (3) in Fig. 16.17. Given a front-end bandwidth of W_{FE} Hz and an in-band interference-to-signal power ratio of P_I/P_S , the resulting effective C/N_0 will be as in (16.8), which for strong interference becomes $C/N_{0,eff} = P_S W_{FE} / P_I$. Thus, to withstand interference exceeding $P_I/P_S = 50$ dB in a $W_{FE} = 10$ MHz bandwidth, a receiver would need to acquire and track GNSS signals below $C/N_{0,eff} = 10 \log_{10}(10^7) - 50 = 20$ dB Hz.

Consumer-grade GNSS receivers offer surprisingly good protection against dense interference despite their low cost, because they have been designed for operation at low C/N_0 . Even without network aiding, a consumer-grade GNSS receiver can acquire signals from a cold start at -148 dBm, which corresponds to $C/N_0 = 26$ dB Hz for a typical $N_0 = -174$ dBm/Hz. This amounts to resilience against P_I/P_S up to 37 dB in a 2 MHz bandwidth. Tracking and performance can be substantially better than cold-start acquisition, achieving remarkable thresholds as

low as -167 dBm, or $C/N_0 = 7$ dB Hz assuming $N_0 = -174$ dBm/Hz [16.78].

The receiver presented in [16.11] can be considered a benchmark for what is possible with a stand-alone scalar-tracking architecture when computational limitations are ignored. Its algorithms can acquire and maintain lock on signals down to $C/N_0 = 18$ dB Hz by assuming a low-cost TCXO and moderate acceleration uncertainty. Clearly, the superior tracking performance of the consumer-grade receiver in [16.78] implies a vectorized tracking architecture.

The current state-of-the-art in low- C/N_0 acquisition and tracking is embodied in the DINGPOS high-sensitivity GNSS platform for deep indoor scenarios [16.79]. The platform records synchronized data from a micro-electromechanical system (MEMS) IMU, a barometer, a magnetometer, and a GNSS RF front-end driven by an OCXO-quality reference clock. The data are combined with known navigation data symbols in a software-defined GNSS receiver employing a vector tracking architecture to achieve coherent integration over 2 s intervals under pedestrian dynamics. In dynamic simulation scenarios, DINGPOS acquires down to $C/N_0 = 6$ dB Hz and tracks down to $C/N_0 = -1$ dB Hz. This represents remarkable interference immunity: up to $P_I/P_S = 71$ dB in a $W_{FE} = 10$ MHz bandwidth for tracking. Even higher P_I/P_S immunity can be achieved by combining DINGPOS-style signal processing with antenna array processing, the subject of the next section.

16.7.3 Antenna-Based Techniques

Though currently expensive, multielement antenna arrays are perhaps the most effective general tool for interference mitigation. Antenna array interference mitigation exploits spatial sparseness in the direction of arrival of interference sources and spatial diversity in the direction of arrival of desired GNSS signals from overhead satellites. Early array processing methods passed the RF signal from each array element through a variable phase shifter. The phase-shifted RF signals were then combined into a single RF stream that was directed to the RF front end for conditioning and digitization. In this approach, the GNSS receiver saw only a single antenna gain pattern (e.g., a pattern with a null directed toward an interference source) at any given instant.

The modern approach to array processing is much more flexible. The RF feed from each antenna is independently digitized, as shown in Fig. 16.17. A complex weight vector is applied across the individual digitized streams to achieve a desired gain pattern. Importantly, any number of weighted combinations of the digital

streams can be created simultaneously, with the unique combinations fed to a bank of separate GNSS processing channels. In this way, each channel sees an alternative antenna array gain pattern, which permits a beam to be steered toward the satellite whose signal the channel is intended to track, for example.

Continuously calculating the set of optimal weighting vectors is the primary computational challenge of array processing, with the primary practical challenge being the need to periodically calibrate the array as temperature and other environmental variations cause minute but significant changes in the phase shift through each antenna element.

A computationally efficient approach to weighting vector calculation is offered in [16.80], but this approach requires the direction of arrival of the desired signal to be known, which entails knowledge of the antenna array's attitude in global coordinates. Preferable are blind adaptive techniques such as the one pro-

posed in [16.81], which automatically maximizes the ratio of power in the desired signal to power in the interference signal plus thermal noise in the correlation products. Better still, though more computationally demanding, are joint space–time interference mitigation techniques that exploit interference time correlation or spatial correlation, or both, in a joint space–time mitigation framework [16.82]. A single interferer is detected in this framework based on estimates of the spatial correlation matrix. A narrowband interferer is detected based on estimates of the time correlation matrix (or based on time correlation evident in the Fourier domain). Such space–time array processing thus combines the virtues of adaptive notch filtering with adaptive beam forming. The beamforming aspect of the approach works equally well whatever the nature of the interference source – intentional or not, GNSS-like or not – so long as the source presents a compact direction of arrival.

References

- 16.1 G.X. Gao, L. Heng, A. Hornbostel, H. Denks, M. Meurer, T. Walter, P. Enge: DME/TACAN interference mitigation for GNSS: Algorithms and flight test results, *GPS Solutions* **17**(4), 561–573 (2013)
- 16.2 Radio Regulations (ITU-R International Telecommunication Union, Radiocommunication Sector, Geneva 2012) <http://www.itu.int/pub/R-REG-RR>
- 16.3 FCC Online Table of Frequency Allocations (Federal Communications Commission, 2013) <http://transition.fcc.gov/oet/spectrum/table/fcctable.pdf>
- 16.4 FCC Enforcement Advisory No. 2011–03, DA 11–249 (Federal Communications Commission, 2011) http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-11-249A1.pdf
- 16.5 Mobile phone and GPS jamming devices FAQ (Australian Communications and Media authority, 2014) <http://www.acma.gov.au/theACMA/faqs-mobile-phone-and-gps-jamming-devices-acma>
- 16.6 G.X. Gao, P. Enge: How many GNSS satellites are too many?, *IEEE Trans. Aerosp. Electron. Syst.* **48**(4), 2865–2874 (2012)
- 16.7 T.E. Humphreys: The GPS dot and its discontents: Privacy vs. GNSS integrity, *Inside GNSS* **7**(2), 44–48 (2012)
- 16.8 M. Rao, C. O'Driscoll, D. Borio, J. Fortuny: Light-squared effects on estimated C/N_0 , pseudoranges and positions, *GPS Solutions* **18**(1), 1–13 (2014)
- 16.9 P. Misra, P. Enge: *Global Positioning System: Signals, Measurements, and Performance*, 2nd edn. (Ganga-Jumana, Lincoln 2012)
- 16.10 A.J. van Dierendonck: GPS receivers. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996) pp. 329–407
- 16.11 M.L. Psiaki, H. Jung: Extended Kalman filter methods for tracking weak GPS signals, *Proc. ION GPS 2002*, Portland (ION, Virginia 2002) pp. 2539–2553, 24–27 Sep. 2002
- 16.12 C. Hegarty, M. Tran, Y. Lee: Simplified techniques for analyzing the effects of non-white interference on GPS receivers, *Proc. ION GPS 2002*, Portland (ION, Virginia 2002) pp. 620–629
- 16.13 J.W. Betz: Effect of narrowband interference on GPS code tracking accuracy, *Proc. ION NTM 2000*, Anaheim (ION, Virginia 2000) pp. 16–27
- 16.14 P.W. Ward, J.W. Betz, C.J. Hegarty: Interference, multipath, and scintillation. In: *Understanding GPS: Principles and Applications*, ed. by E.D. Kaplan, C.J. Hegarty (Artech House, Boston 2005) pp. 243–299
- 16.15 F.M. Gardner: *Phaselock Techniques*, 3rd edn. (Wiley, Hoboken 2005)
- 16.16 T.E. Humphreys, M.L. Psiaki, B.M. Ledvina, P.M. Kintner Jr: GPS carrier tracking loop performance in the presence of ionospheric scintillations, *Proc. ION GNSS 2005*, Long Beach (ION, Virginia 2005) pp. 156–167
- 16.17 M.K. Simon, M. Alouini: *Digital Communications over Fading Channels* (Wiley, New York 2000)
- 16.18 A.J. Viterbi: *Principles of Coherent Communication* (McGraw-Hill, New York 1966)
- 16.19 S.C. Gupta: Phase-locked loops, *Proc. IEEE* **63**(2), 291–306 (1975)
- 16.20 G. Ascheid, H. Meyr: Cycle slips in phase-locked loops: A tutorial survey, *IEEE Trans. Comm.* **COM-30**(10), 2228–2241 (1982)
- 16.21 B. Motella, S. Savasta, D. Margaria, F. Dovis: Method for assessing the interference impact on GNSS receivers, *IEEE Trans. Aerosp. Electron. Syst.* **47**(2),

- 1416–1432 (2011)
- 16.22 J.J. Spilker Jr.: GPS signal structure and theoretical performance. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996) pp. 57–119
 - 16.23 Navstar GPS Space Segment / Navigation User Segment Interfaces, Interface Specification, IS-GPS-200, Rev. H (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo 2013)
 - 16.24 J.J. Spilker Jr.: Interference effects and mitigation techniques. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996) pp. 717–771
 - 16.25 J.H. van Vleck, D. Middleton: The spectrum of clipped noise, *Proc. IEEE* **54**(1), 2–19 (1966)
 - 16.26 F. Amoroso: Adaptive A/D converter to suppress CW interference in DSPN spread-spectrum communications, *IEEE Trans. Comm.* **31**, 1117–1123 (1983)
 - 16.27 C.J. Hegarty: Analytical model for GNSS receiver implementation losses, *Navigation* **58**(1), 29–44 (2011)
 - 16.28 J. Max: Quantizing for minimum distortion, *IRE Trans. Inf. Theory* **6**(1), 7–12 (1960)
 - 16.29 F. Amoroso, J.L. Bricker: Performance of the adaptive A/D converter in combined CW and Gaussian interference, *IEEE Trans. Comm.* **34**(3), 209–213 (1986)
 - 16.30 D.J. McLean, N.R. Labrum: *Solar Radiophysics: Studies of Emission from the Sun at Metre Wavelengths* (Cambridge Univ. Press, New York 1985)
 - 16.31 P.M. Kintner Jr., T.E. Humphreys, J. Hinks: GNSS and ionospheric scintillation: How to survive the NextSolar maximum, *Inside GNSS* **4**(4), 22–30 (2009)
 - 16.32 R.V. Jones: *Most Secret War* (Penguin UK, London 2009)
 - 16.33 A.P. Cerruti, P.M. Kintner, D.E. Gary, A.J. Mannucci, R.F. Meyer, P. Doherty, A.J. Coster: Effect of intense December 2006 solar radio bursts on GPS receivers, *Space Weather* **6**(10), 1–10 (2008)
 - 16.34 A.P. Cerruti, P.M. Kintner, D.E. Gary, L.J. Lanzerotti, E.R. de Paula, H.B. Vo: Observed solar radio burst effects on GPS/wide area augmentation system carrier-to-noise ratio, *Space Weather* **4**(10), 1–9 (2006)
 - 16.35 D.M. Akos: Who's afraid of the spoofer? GPS/GNSS spoofing detection via automatic gain control (AGC), *Navigation* **59**(4), 281–290 (2012)
 - 16.36 G.M. Nita, D.E. Gary, L.J. Lanzerotti, D.J. Thomson: The peak flux distribution of solar radio bursts, *Astrophys. J.* **570**, 423–438 (2002)
 - 16.37 T.E. Humphreys, M.L. Psiaki, B.M. Ledvina, A.P. Cerruti, P.M. Kintner: A data-driven testbed for evaluating GPS carrier tracking loops in ionospheric scintillation, *IEEE Trans. Aerosp. Electron. Syst.* **46**(4), 1609–1623 (2010)
 - 16.38 T.E. Humphreys, M.L. Psiaki, P.M. Kintner: Modeling the effects of ionospheric scintillation on GPS carrier phase tracking, *IEEE Trans. Aerosp. Electron. Syst.* **46**(4), 1624–1637 (2010)
 - 16.39 J. Aarons: Global morphology of ionospheric scintillations, *Proc. IEEE* **70**(4), 360–378 (1982)
 - 16.40 J. Aarons: Global positioning system phase fluctuations at auroral latitudes, *J. Geophys. Res.* **102**, 17219–17231 (1997)
 - 16.41 B.M. Ledvina, J.J. Makela, P.M. Kintner: First observations of intense GPS L1 amplitude scintillations at midlatitude, *Geophys. Res. Lett.* **29**(14), 4–1–4–4 (2002)
 - 16.42 T.E. Humphreys, M.L. Psiaki, J.C. Hinks, B. O'Hanlon, P.M. Kintner Jr.: Simulating ionosphere-induced scintillation for testing GPS receiver phase tracking loops, *IEEE J. Sel. Top. Signal Process.* **3**(4), 707–715 (2009)
 - 16.43 J. Do, D.M. Akos, P.K. Enge: L and S bands spectrum survey in the San Francisco Bay area, *Proc. IEEE PLANS 2004, Monterey* (IEEE, New York 2004) pp. 566–572
 - 16.44 R. Johannessen, S.J. Gale, M.J.A. Asbury: Potential interference sources to GPS and solutions appropriate for applications to civil aviation, *IEEE Aerosp. Electron. Syst. Mag.* **5**(1), 3–9 (1990)
 - 16.45 A.T. Balaei, A.G. Dempster: A statistical interference technique for GPS interference detection, *IEEE Trans. Aerosp. Electron. Syst.* **45**(4), 1499–1511 (2009)
 - 16.46 C. Kurby, R. Lee, L. Cygan, E. Derbez: Maintaining precision receiver performance while rejecting adjacent band interference, *Proc. ION ITM 2012, Newport Beach* (ION, Virginia 2012) pp. 574–597
 - 16.47 R.H. Mitch, R.C. Dougherty, M.L. Psiaki, S.P. Powell, B.W. O'Hanlon, J.A. Bhatti, T.E. Humphreys: Signal characteristics of civil GPS jammers, *Proc. ION GNSS 2011, Portland* (ION, Virginia 2011) pp. 1907–1919
 - 16.48 K.D. Wesson, T.E. Humphreys: Hacking drones, *Sci. Am.* **309**(5), 54–59 (2013)
 - 16.49 R.H. Mitch, M.L. Psiaki, S.P. Powell, B.W. O'Hanlon: Signal acquisition and tracking of chirp-style GPS jammers, *Proc. ION GNSS 2013, Nashville* (ION, Virginia 2013) pp. 2893–2909
 - 16.50 A.J. Kerns, D.P. Shepard, J.A. Bhatti, T.E. Humphreys: Unmanned aircraft capture and control via GPS spoofing, *J. Field Robotics* **31**(4), 617–636 (2014)
 - 16.51 J. Bhatti: Sensor Deception Detection and Radio-Frequency Emitter Localization, Ph.D. Thesis (University of Texas, Austin 2015)
 - 16.52 European GNSS (Galileo) Open Service Signal In Space Interface Control Document, OS SIS ICD, Iss. 1.2 (European Union 2015)
 - 16.53 L. Scott: Anti-spoofing and authenticated signal architectures for civil navigation systems, *Proc. ION GNSS 2003, Portland* (ION, Virginia 2003) pp. 1542–1552
 - 16.54 T.E. Humphreys: Detection strategy for cryptographic GNSS anti-spoofing, *IEEE Trans. Aerosp. Electron. Syst.* **49**(2), 1073–1090 (2013)
 - 16.55 K.D. Wesson, M.P. Rothlisberger, T.E. Humphreys: Practical cryptographic civil GPS signal authentication, *Navigation* **59**(3), 177–193 (2012)
 - 16.56 A.J. Kerns, K.D. Wesson, T.E. Humphreys: A blueprint for civil GPS navigation message authentication, *Proc. IEEE/ION PLANS 2014, Monterey* (ION, Virginia 2014) pp. 262–269
 - 16.57 I. Fernandez Hernandez, V. Rijmen, G. Seco Grana-dos, J. Simon, I. Rodriguez, J.D. Calle: Design drivers, solutions and robustness assessment of navigation message authentication for the Galileo

- open service, Proc. ION GNSS 2014, Tampla (ION, Virginia 2014) pp. 2810–2827
- 16.58 T.E. Humphreys, B.M. Ledvina, M.L. Psiaki, B.W. O'Hanlon, P.M. Kintner Jr.: Assessing the spoofing threat: Development of a portable GPS civilian spoofer, Proc. ION GNSS 2008, Savannah (ION, Virginia 2008) pp. 2314–2325
 - 16.59 G. Hein, F. Kneissl, J.-A. Avila-Rodriguez, S. Wallner: Authenticating GNSS: Proofs against spoofs, Part 2, Inside GNSS 2(5), 71–78 (2007)
 - 16.60 Vulnerability assessment of the transportation infrastructure relying on the Global Positioning System (John A. Volpe National Transportation Systems Center 2001)
 - 16.61 O. Pozzobon, C. Wullems, M. Detratti: Security considerations in the design of tamper resistant GNSS receivers, Proc. NAVITEC 2010, Noordwijk (IEEE, New York 2010)
 - 16.62 T.E. Humphreys, D.P. Shepard, J.A. Bhatti, K.D. Wesson: A testbed for developing and evaluating GNSS signal authentication techniques, Proc. Int. Symp. Certif. GNSS Syst. Serv. (CERGAL), Dresden (DGON, Düsseldorf 2014)
 - 16.63 K. Wesson: Secure Navigation and Timing Without Local Storage of Secret Keys, Ph.D. Thesis (University of Texas, Austin 2015)
 - 16.64 V. Dehghanian, J. Nielsen, G. Lachapelle: GNSS spoofing detection based on receiver C/N_0 estimates, Proc. ION GNSS 2012, Nashville (ION, Virginia 2012) pp. 2878–2884
 - 16.65 P.W. Ward: GPS receiver RF interference monitoring, mitigation, and analysis techniques, Navigation 41(4), 367–391 (1994)
 - 16.66 S. Lo, D. Akos, F.M. Eklof, O. Isoz, H. Borowski: Detecting false signals with automatic gain control, GPS World 23(4), 38–43 (2012)
 - 16.67 T.E. Humphreys, J.A. Bhatti, D.P. Shepard, K.D. Wesson: The texas spoofing test battery: Toward a standard for evaluating GNSS signal authentication techniques, Proc. ION GNSS 2012, Nashville (ION, Virginia 2012) pp. 3569–3583
 - 16.68 National Oceanic and Atmospheric Administration: Space Weather Alerts Description and Criteria, <http://legacy-www.swpc.noaa.gov/alerts/description.html#electron>
 - 16.69 A. Jafarnia-Jahromi, A. Broumandan, J. Nielsen, G. Lachapelle: Pre-despreading authenticity verification for GPS L1 C/A signals, Navigation 61(1), 1–11 (2014)
 - 16.70 S. Gunawardena, Z. Zhu, M.U. de Haag, F. van Graas: Remote-controlled, continuously operating GPS anomalous event monitor, Navigation 56(2), 97–113 (2009)
 - 16.71 A.S. Willsky: A survey of design methods for failure detection in dynamic systems, Automatica 12(6), 601–611 (1976)
 - 16.72 M.L. Psiaki, B.W. O'Hanlon, J.A. Bhatti, D.P. Shepard, T.E. Humphreys: GPS spoofing detection via dual-receiver correlation of military signals, IEEE Trans. Aerosp. Electron. Syst. 49(4), 2250–2267 (2013)
 - 16.73 M.L. Psiaki, S.P. Powell, B.W. O'Hanlon: GNSS spoofing detection using high-frequency antenna motion and carrier-phase data, Proc. ION GNSS 2013, Nashville (ION, Virginia 2013) pp. 2949–2991
 - 16.74 M.L. Psiaki, B.W. O'Hanlon, S.P. Powell, J.A. Bhatti, K.D. Wesson, T.E. Humphreys, A. Schofield: GNSS spoofing detection using two-antenna differential carrier phase, Proc. ION GNSS 2014, Tampa (ION, Virginia 2014) pp. 2776–2800
 - 16.75 S. Khanafseh, N. Roshan, S. Langel, F.-C. Chan, M. Joerger, B. Pervan: GPS spoofing detection using RAIM with INS coupling, Proc. IEEE/ION PLANS 2014, Monterey (ION, Virginia 2014) pp. 1232–1239
 - 16.76 L.B. Milstein: Interference rejection techniques in spread spectrum communications, Proc. IEEE 76(6), 657–671 (1988)
 - 16.77 G. Dimos, T.N. Upadhyay, T. Jenkins: Low-cost solution to narrowband GPS interference problem, Proc. Nat. Aerosp. Electron. Conf. NAECON 1995, Dayton (IEEE, New York 1995) pp. 145–153
 - 16.78 MAX-M8 GNSS Module datasheet (u-Blox), UBX-15031506, [https://www.u-blox.com/sites/default/files/MAX-M8-FW3_DataSheet_\(UBX-15031506\).pdf](https://www.u-blox.com/sites/default/files/MAX-M8-FW3_DataSheet_(UBX-15031506).pdf)
 - 16.79 H. Niedermeier, B. Eissfeller, J. Winkel, T. Pany, B. Riedl, T. Wörz, R. Schweikert, S. Lagrasta, G. Lopez-Risueno, D. Jimenez-Banos: DINGPOS: High sensitivity GNSS platform for deep indoor scenarios, Proc. Int. Conf. Indoor Position. Indoor Navig. (IPIN), Zurich (IEEE, New York 2010)
 - 16.80 G. Seco-Granados, J.A. Fernández-Rubio, C. Fernández-Prades: ML estimator and hybrid beamformer for multipath and interference mitigation in GNSS receivers, IEEE Trans. Signal Process. 53(3), 1194–1208 (2005)
 - 16.81 M. Sgammini, F. Antreich, L. Kurz, M. Meurer, T.G. Noll: Blind adaptive beamformer based on orthogonal projections for GNSS, Proc. ION GNSS 2012, Nashville (ION, Virginia 2012) pp. 926–935
 - 16.82 M.H. Castañeda, M. Stein, F. Antreich, E. Tasdemir, L. Kurz, T.G. Noll, J.A. Nassek: Joint space-time interference mitigation for embedded multi-antenna GNSS receivers, Proc. ION GNSS 2013, Nashville (ION, Virginia 2013) pp. 3399–3408

Antennas

17. Antennas

Moazam Maqsood, Steven Gao, Oliver Montenbruck

The basic purpose of a global navigation satellite system (GNSS) user antenna is the reception of navigation signals from all visible GNSS satellites. Transmit antennas onboard the GNSS satellites, on the other hand, are quite different and employ large antenna arrays to create high-gain global beams illuminating the entire surface of the Earth.

This chapter presents different design options for GNSS antennas operating in the L-band of the radio frequency spectrum. It starts with a brief discussion of key requirements for the GNSS receiving antenna, where several design parameters are introduced and explained. Thereafter, antennas of different design technologies suitable to GNSS are explored and discussed in detail. Following the introduction of major antenna candidates, different variants for specialized requirements, such as the small form factor or multipath mitigation are presented. Complementary to receiving antennas, the design of antenna arrays for signal transmission on the GNSS satellites is presented next, along with a discussion on specific antennas employed on the Global Positioning System (GPS), Galileo, Global'naya Navigatsionnaya Sputnikova Sistema (GLONASS) and BeiDou satellites. Finally, a comprehensive discussion on antenna measurements and the performance evaluation is provided.

17.1	GNSS Antenna Characteristics	506	17.1.7	Axial Ratio	508
17.1.1	Center Frequency	507	17.1.8	Impedance Matching and Return Loss	508
17.1.2	Bandwidth	507	17.1.9	Front-to-Back and Multipath Ratio ...	509
17.1.3	Radiation Pattern	507	17.1.10	Phase-Center Stability	509
17.1.4	Antenna Gain	507	17.2	Basic GNSS Antenna Types	509
17.1.5	3 dB Beam Width	507	17.2.1	Microstrip Patch Antenna	509
17.1.6	Polarization	508	17.2.2	Helix Antenna.....	510
			17.2.3	Quadrifilar Helix Antenna	511
			17.2.4	Spiral Antenna	512
			17.2.5	Wide-Band Bow-Tie	513
			17.2.6	Wide-Band Pinwheel Antenna	513
			17.3	Application-Specific GNSS Antennas ..	513
			17.3.1	Hand-Held Terminals	513
			17.3.2	Surveying and Geodesy.....	514
			17.3.3	Aviation	515
			17.3.4	Space Applications	516
			17.3.5	Antijamming Antennas.....	517
			17.3.6	GNSS Remote Sensing	518
			17.4	Multipath Mitigation	519
			17.4.1	Metallic Reflector Ground Plane.....	520
			17.4.2	Choke-Ring Ground Plane.....	520
			17.4.3	Noncutoff Corrugated Ground Plane...	521
			17.4.4	Convex Impedance Ground Plane	521
			17.4.5	3-D Choke-Ring Ground Plane	521
			17.4.6	Cross Plate Reflector Ground Plane.....	522
			17.4.7	Electromagnetic Band Gap (EBG) Substrate	522
			17.5	Antennas for GNSS Satellites	523
			17.5.1	Concentric Helix Antenna Arrays	523
			17.5.2	Patch Antenna Arrays.....	524
			17.5.3	Reflector-Backed Monofilar Antenna .	526
			17.6	Antenna Measurement and Calibration	527
			17.6.1	Basic Antenna Testing.....	527
			17.6.2	Phase-Center Calibration	528
			References	531	

17.1 GNSS Antenna Characteristics

Antennas are basic elements of any radio frequency (RF) system, be it used for audio and video broadcasting, communication, or radio navigation. An antenna serves as an interface between the electric circuitry of the RF system and free space [17.1, 2]. In a transmit antenna, electric currents are converted into electro-magnetic waves. On the contrast, a receiver antenna converts those radio waves back into electric currents. However, any antenna can, in principle, serve both purposes. The distinction is thus largely driven by the practical application, since specific antenna aspects may be optimized for either of the two functions.

Global navigation satellite systems (GNSSs) such as the US global positioning system (GPS), the Russian Global'naya Navigatsionnaya Sputnikova Sistema (GLONASS), the European Galileo, and the Chinese BeiDou transmit their signals in the so-called L-band of the RF spectrum. Within this band, which covers a total frequency range of 1–2 GHz, various types of communication services have been allocated by the International Telecommunications Union (ITU), but frequencies of 1164–1300 and 1559–1610 MHz have been specifically assigned to radio navigation satellite services (RNSSs). With the exception of the Indian regional navigation satellite system (IRNSS) and the third-generation BeiDou system, that expanded into the S-band and use a frequency near 2.2 GHz for selected navigation signals, all global and regional navigation systems as well as satellite-based augmentation systems (SBAS) are presently confined to the aforementioned L-band frequencies.

Figure 17.1 illustrates the frequency allocation chart for individual navigation satellite systems. The close (and in some cases overlapping) frequency allocation has both its advantages and disadvantages. The advantage is that all the designated frequencies for GNSS lie in a close vicinity, thereby making it possible for a single wide-band antenna to receive signals from several systems. Moreover, the use of re-configurable antennas

and RF front ends becomes possible. On the other hand, the potential disadvantage of such an arrangement is possible interference from one system into another. The interesting fact is that all navigation satellite systems use different signal formatting and modulation forms, thereby providing natural immunity against the interference due to frequency overlapping (Chap. 4).

In order to understand the design and performance aspects of a GNSS antenna, we will start with the requirements of an ordinary GPS antenna [17.3]. This approach is realistic as the GPS is a widely used navigation system. However, the same discussion, with a few modifications, can be applied to other systems as well. Table 17.1 presents representative design requirements of a GPS antenna suitable for receiving all signals transmitted by this navigation system. An important point to note is that these are only basic requirements of a GNSS antenna design. Some applications may also put a set of particular requirements such as phase-center (PC) stability, cross-polarization isolation, axial ratio (AR), and and so on. A brief overview of these requirements is presented in the following sections for the reader to appreciate the challenges these requirements can put on the antenna engineer.

Table 17.1 Example requirements of a triple-frequency GPS antenna for civil users

Parameter	Value
Center frequency	L1: 1575.420 L2: 1227.600 L5: 1176.450
Bandwidth	L1: $\approx 2 \times 10.023$ MHz L2: $\approx 2 \times 10.23$ MHz L5: $\approx 2 \times 10.23$ MHz
Radiation pattern	Semi-hemispherical
Gain	$\approx 3\text{--}5$ dBiC
Polarization	Right-hand circular polarization
Return loss	< -10 dB
Impedance	50 Ω

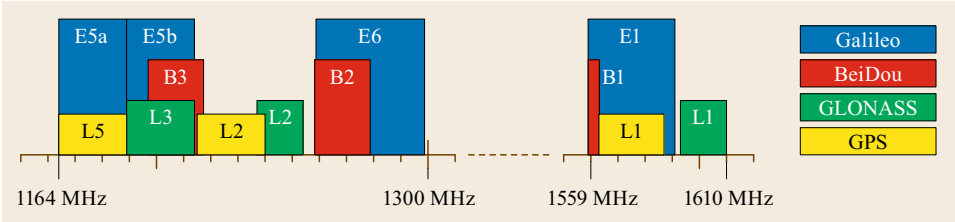


Fig. 17.1 Frequency allocation for navigation satellite systems in the lower and upper L-band range assigned for radio navigation satellite services. Frequencies for BeiDou refer to the second generation system (BDS-2) which started its regional service in 2012

17.1.1 Center Frequency

Center frequency or the frequency of operation is the value at which the complete RF system including the antenna is designed. In the case of GPS, three center frequencies are mentioned (L1, L2, and L5). The L1 frequency (1.575 GHz) is the primary GPS frequency and has been designated for the open standard positioning service (SPS). The precise positioning service (PPS), in contrast, makes the joint use of both the L1 and L2 frequencies to enable ionospheric corrections, and similar options have become available for general users with the addition of the new civil L2C signal on the L2 frequency. Finally, a new L5 signal has been introduced, which is mainly targeted at aviation users. To support all related applications, a modern GPS antenna must, as a minimum, cover the L1, L2, and L5 frequencies. An even wider set of center frequencies must be covered by a generic multi-GNSS antenna.

17.1.2 Bandwidth

The bandwidth of an antenna is normally defined as the range of frequencies within which the antenna operates successfully (i.e., fulfilling the entire design requirements). The term bandwidth can be further classified into the impedance bandwidth and the gain bandwidth. In terms of a GNSS antenna, another important consideration is whether the antenna maintains the required right-hand circular polarization (RHCP) within the given bandwidth. A common bandwidth of ± 10.23 MHz has been required in Table 17.1 to cover all signals transmitted by the GPS satellites at the respective center frequencies. Antennas designed to support only the C/A-code and L2C signals might, however, be confined to a ± 1.023 MHz bandwidth in view of the 10-times lower chipping rate of these signals.

17.1.3 Radiation Pattern

The *radiation pattern* of an antenna is a representation of radiation properties such as the electric field or power as a function of spatial coordinates. These patterns can be represented in two- (2-D) and three-dimensional (3-D) coordinates, and are commonly presented as functions of observation angles around the antenna. A trace of the received electric field at a constant radius is called the amplitude field pattern, while a graph of the spatial variation of the power density along a constant radius is called an amplitude power pattern. Often the power pattern is plotted on a logarithmic scale or more commonly in decibels (dB). This is done to achieve a more detailed representation of areas with low power. In the case of a GNSS receiving

antenna, the antenna should have a semi-hemispherical pattern (directed toward the transmitting satellite). On the other hand, the radiation pattern of a transmitting GNSS antenna is a sharp directional beam compensating for the notable free-space loss.

17.1.4 Antenna Gain

The gain G of an antenna [17.2] describes, how well an antenna can receive power from or transmit into a specific direction in comparison to an idealized, lossless isotropic antenna. It is defined as the product $G = eD$ of the efficiency e of energy conversion inside the antenna and the directivity

$$D = \frac{4\pi U}{P_{\text{rad}}}, \quad (17.1)$$

where U is the antenna's radiation intensity (power per solid angle) in a given direction and $P_{\text{rad}}/(4\pi)$ is the corresponding value (ratio of radiated energy per solid angle of a sphere) for the isotropic radiator. The gain is commonly expressed in a logarithmic scale

$$G_{\text{dBi}} = 10 \log \left(\frac{G}{10} \right), \quad (17.2)$$

where dBi denotes decibels relative to an isotropic antenna. More specifically, dBiC is used for gain specifications when working with circular polarization. If no direction is specified, gain values are usually referred to the boresight direction (i.e., the main symmetry axis of the gain pattern). While GNSS receiving antennas are typically low-gain antennas enabling reception over a large range of angles, highly directive (high-gain) antennas are used for transmitting signals on the GNSS satellites.

17.1.5 3 dB Beam Width

The antenna beam width is normally used to define the area of maximum power concentration. Graphically, it is represented by the angular separation between two identical points around the antenna maximum. Two most common methods of representing an antenna beam width are the half-power beam width (HPBW) and the first-null beam width (FNBW). When plotted in decibels, the HPBW is also referred to as the 3 dB beam width. In the case of the GNSS receiving antenna, the required antenna beam should be as wide as possible. This ensures maximum satellite visibility. On the other hand, the transmitting satellite shall have a narrow beam width as it requires more focus into a particular direction. Figure 17.2 shows the radiation pattern of a GNSS

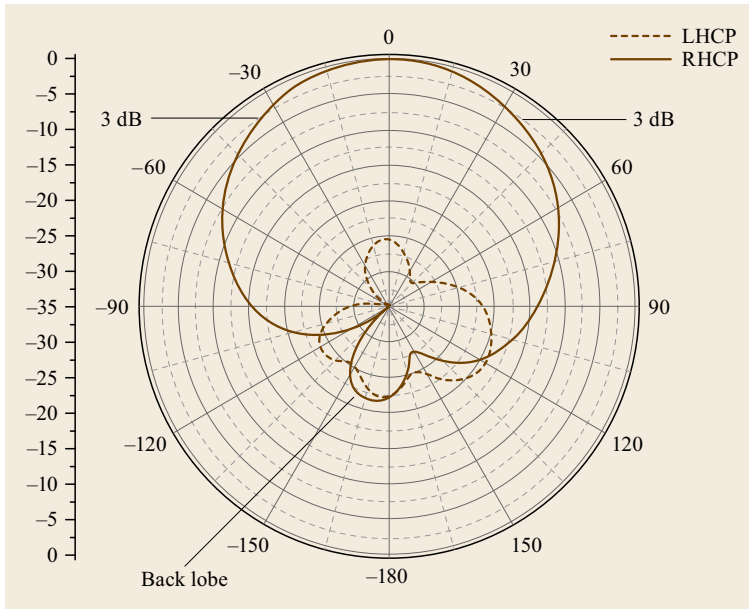


Fig. 17.2 Simulated radiation pattern of a GNSS antenna. Values are normalized relative to the boresight direction

antenna, where a dominant semi-hemispherical pattern can be seen for the RHCP. The 3 dB beam width points have also been labeled.

17.1.6 Polarization

When an antenna emits electromagnetic waves, it has associated electric and magnetic field vectors. The polarization of any antenna is generally defined as the orientation of its electric (E)-field vector. If the E-field vector is aligned with the horizon, the antenna is said to be horizontally polarized, whereas for the E-field vector to be aligned perpendicular to the horizon, the antenna is said to be vertically polarized. In the case of GNSS, the conventional polarization is neither horizontal nor vertical, but RHCP. This means that the electric field of a GNSS signal is composed of two orthogonal waves of equal amplitude with a 90° phase shift. The resultant E-field then circulates in a clockwise direction. Upon reflection, the polarization may change, thus giving rise to left-hand circularly polarized (LHCP) waves. Ideally, such signal components should be fully suppressed by a GNSS receiving antenna.

17.1.7 Axial Ratio

The AR is the ratio of the magnitudes of the major and minor axes of the polarization ellipse. It is equal to 1 for purely circular polarization and grows with increasing ellipticity. For linear polarization, the AR is infinite, because one of the orthogonal components of the field is zero.

An ideal GNSS antenna would thus exhibit an AR of 0 dB, but a value of < 3 dB is generally acceptable. Since the AR tends to increase with increasing boresight angle, it is common to indicate the range of boresight angles for which this condition is met. The 3 dB AR beam width is a key parameter for advocating the performance of circularly polarized antennas. The wider this beam width, the greater the ability of the antenna to reject multipath signals.

17.1.8 Impedance Matching and Return Loss

Impedance matching and return loss are the two parameters that tell an antenna engineer, how well an antenna will accept the incoming power. Regardless of the antenna type and the frequency of operation, it is essential for the antenna to have a good impedance matching with the feed line and a high return loss. As mentioned in Table 17.1, the input impedance requirement for a GNSS antenna is 50Ω , a common standard, which makes the antennas of different manufacturers compatible with the RF system. The return loss describes the ratio between the power fed into a transmit antenna and the power reflected back into the feed cable. It is usually required to be less than -10 dB, which ensures that at least 90% of the incoming power is transmitted to the antenna for radiation.

As an alternative to the return loss, the reflection coefficient $|S_{11}|$ and the voltage standing wave ratio (VSWR) are occasionally provided in antenna specifications. They relate to the ratio of the reflected and incoming field amplitude and can be converted to return

loss by the relations

$$\text{return loss} = -20 \log |s_{11}| = -20 \log \left| \frac{\text{VSWR} - 1}{\text{VSWR} + 1} \right|. \quad (17.3)$$

17.1.9 Front-to-Back and Multipath Ratio

The front-to-back ratio (**FBR**) relates the antenna energy directed in the boresight (main lobe) direction to that of the backlobes

$$\text{FBR} = \frac{G(\theta = 0^\circ)}{G(\theta = 180^\circ)}. \quad (17.4)$$

The ratio simultaneously highlights an antenna's directivity and resistance to multipath. Generally, a high FBR is required for a GNSS antenna, since it will allow the antenna to attenuate signals (e.g., ground reflections) coming from unwanted directions. The FBR is influenced by a combination of the antenna's backside shielding and the sensitivity to LHCP signals.

A related parameter, known as the multipath rejection ratio (**MPR**)

$$\text{MPR} = \frac{G_{\text{RHCP}}(\theta)}{G_{\text{LHCP}}(180^\circ - \theta) + G_{\text{RHCP}}(180^\circ - \theta)}, \quad (17.5)$$

17.2 Basic GNSS Antenna Types

GNSS antennas can be categorized in more than one ways: based on the associated design technology such as microstrip, helix, spiral, and so on, or according to the subject application such as navigation, surveying, remote sensing, and antijamming. This section takes the former approach and introduces the basic GNSS antenna candidates along with a discussion of their design methodology. It is complemented later by a dedicated section on application-specific GNSS antennas (Sect. 17.3), which presents several variants of the basic antenna types for specialized requirements.

The two most common and popular GNSS antenna configurations are the printed (microstrip patch) and helix antennas [17.5, 6]. Both of these antenna types can provide a reasonably small size and are widely used for single-frequency, mass market applications. Even though the concepts may be extended to dual- or even multiband antennas, a much simpler option is to use a wide-band antenna covering almost the entire L-band. However, such an antenna will require sharp filtering to isolate adjacent frequency bands and the associated receiver should have a mechanism in place

directly compares the gain for RHCP signals from a given boresight angle θ with those for RHCP and LHCP radiation received from $180^\circ - \theta$ [17.4]. For a zenith-looking antenna, this angle corresponds to the incidence angle of ground-reflected signals. The MPR is thus an important indicator for characterizing the multipath mitigation performance of an antenna.

17.1.10 Phase-Center Stability

Conceptually, the antenna phase-center is a point within the antenna radiation pattern, where all the power emanates from (for a TX antenna) or converges to (for a RX antenna). This point is typically different from the geometric center of an antenna and also depends on the signal frequency. In practice, the wavefronts of an antenna will also deviate from the ideal case of the concentric spherical shell. These deviations give rise to direction-dependent delays in GNSS measurements, which are known as phase-center variations (**PCVs**). Accurate knowledge of the phase-center offset (**PCO**) and PCV from an established mechanical reference point is most important for precise positioning applications and requires proper calibration for both GNSS satellite (transmit) and user (receive) antennas. Ensuring a good PC stability is generally most challenging for wide- or multiband antennas.

to reject out-of-band interference coming, for example, from amateur radio and digital audio–video broadcast (**DVB**, **DAB**) services.

17.2.1 Microstrip Patch Antenna

The microstrip patch (or simply patch) is a simple antenna configuration, consisting of a metallic patch that is mounted above a conducting ground plane with a dielectric substrate separating both layers (Fig. 17.3, [17.7]). The top layer is designed so that it resonates at a particular frequency, thereby acting as a radiating antenna, whereas the bottom layer acts as a ground plane and is necessary for successful antenna operation. Microstrip antennas have the advantages of low profile, small form factor, surface conformability, and moderate gain performance, while its disadvantages include narrow bandwidth and emission of surface waves. Despite these disadvantages, the microstrip antenna is the most popular choice for GNSS antennas and is particularly attractive for mass market applications.

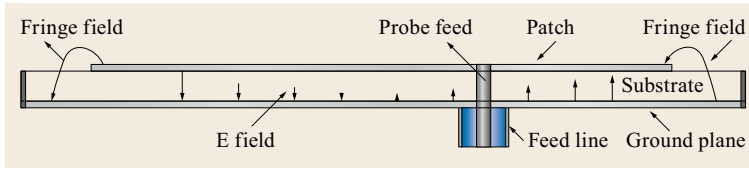


Fig. 17.3 Side view of a patch antenna with fringing field

Even though different shapes may be considered for general patch antennas, a square patch is commonly employed for antennas for the reception or transmission of RF signals with circular polarization. Together, the patch and ground plane form a cavity with a resonant frequency f_r that depends on the width w of the patch element and the permittivity ϵ_r of the substrate. As a rule of thumb, resonance is achieved if the patch dimension equals half the signal wavelength in the dielectric material, that is,

$$w = \frac{c}{2f_r \sqrt{\epsilon_r}}, \quad (17.6)$$

with c denoting the speed of light. While a substrate of low permittivity ($\epsilon_r \approx 1$) would result in a GNSS antenna size of about 1 dm, substantially more compact patch antennas can be designed with ceramic materials offering a 10–100 times higher permittivity.

As illustrated in Fig. 17.3, the electric field lines connecting the patch and the ground plane pass through two different media (air and substrate). This creates a fringing effect, which increases the *electronic width* of the microstrip line. Even though the field lines are mostly concentrated under the patch itself, whenever the width of the patch is much larger than the substrate thickness h , the fringing affects the resonance frequency and must be compensated by a corresponding length correction. In order to achieve a resonance at the desired frequency f_r , a slightly smaller patch size

$$w' = w - 2\Delta w \quad (17.7)$$

is required due to the fringing than given by the simplified relation (17.6). Following [17.8], the correction can be approximated by the expression

$$\Delta w = 0.412 h \frac{(\epsilon_{\text{eff}} + 0.3) \left(\frac{w}{h} + 0.264 \right)}{(\epsilon_{\text{eff}} - 0.258) \left(\frac{w}{h} + 0.8 \right)}, \quad (17.8)$$

where

$$\epsilon_{\text{eff}} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \left(1 + 10 \frac{h}{w} \right)^{-\frac{1}{2}}, \quad (17.9)$$

denotes an effective permittivity.

One of the biggest advantages of a patch antenna is that it can be designed with more than one shape.

Rectangular, circular, diamond, and ring are the few common shapes that can be employed for patches. However, with reference to GNSS, the shape of the patch needs to be symmetric (square or circular) as the achievement of circular polarization is only possible if the patch produces similar but orthogonal electric fields. This is achieved by feeding the patch antenna at two orthogonal points and then combining them by means of a quadrature hybrid coupler or a delay line mechanism (Fig. 17.4a). An alternative design approach is to use a single-feed square patch antenna with two truncated corners on opposite sides (Fig. 17.4b), where the position of the feed relative to the corners determines the rotation of the circular polarization. This approach is very simple, easy to implement, and results in an inexpensive antenna. However, the achievement of good polarization purity requires rigorous tuning of the corners.

Microstrip patch antennas can also be used for multiband operation by stacking more than one patch elements over one another [17.9, 10], where different stacked layers are fed together using proximity or aperture coupling.

17.2.2 Helix Antenna

Next to the microstrip patch antenna, the helix antenna [17.1, 9] represents another basic antenna type for GNSS applications. The antenna consists of a wire shaped in the form of a helix with the longitudinal axis perpendicular to the ground plane (Fig. 17.5).

The radiation pattern of the helix antenna can be controlled by varying geometrical parameters of the antenna. The same is true for the input impedance of the antenna, which primarily depends upon the pitch angle and size of the conducting wire. The general polariza-

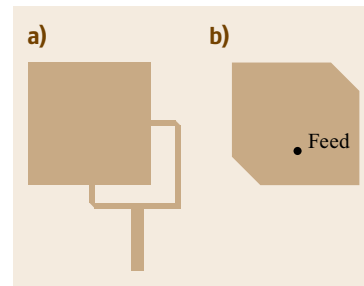


Fig. 17.4a,b Patch antenna with feed network for circular polarization (a) and patch antenna with truncated corners (b)

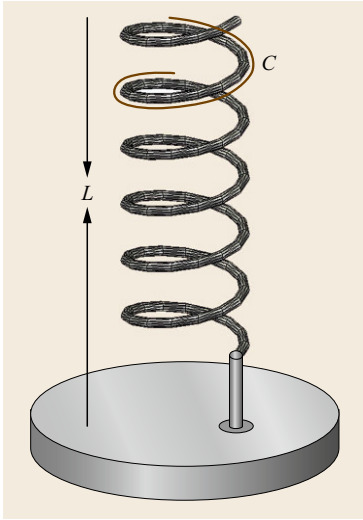


Fig. 17.5 Geometrical parameters of a GNSS helix antenna

tion of the antenna is elliptical as it has both vertical and horizontal sections. While the antenna acts like a dipole with an almost omnidirectional gain pattern and linear polarization if its dimension is small compared to the wavelength (*broadside* or *normal* mode), it attains a high directivity (*axial* or *end-fire* mode) and circular polarization in the opposite case.

In reference to GNSS, the helix is normally operated in the axial (end-fire) mode to obtain a semi-hemispherical pattern. For circular polarization in the main lobe, the number of turns should be greater than 3, which implies that the antenna size may not be suitable for use in handheld devices. However, the helix antenna is a very popular candidate for GNSS satellites, where it is used for transmitting the navigation signals in a high-gain beam. The helix normally operates with a ground plane, whether flat or cupped in the shape of a cylindrical cavity. The typical diameter of the ground plane amounts to three quarters of the signal wavelength.

To operate the helix in the axial mode with circular polarization, the circumference (C), pitch angle (α), and number of turns N of the helix should remain within the following limits

$$\begin{aligned} \frac{3}{4} < \frac{C}{\lambda} < \frac{4}{3}, \\ 12^\circ < \alpha < 14^\circ, \\ 3 < N. \end{aligned} \quad (17.10)$$

Once the above parameters are fixed, the spacing between the turns S and the antenna length L can be obtained as

$$S = C \tan \alpha, \quad L = NS. \quad (17.11)$$

With the above quantities, key parameters of the helix antenna such as the input resistance R , the directivity D , and the HPBW can be obtained from the approximate relations

$$\begin{aligned} R &= 140 \frac{C}{\lambda}, \quad D = 15N \frac{C^2 S}{\lambda^3}, \\ \text{HPBW} &= 52^\circ \frac{\lambda^{\frac{3}{2}}}{C \sqrt{NS}}, \end{aligned} \quad (17.12)$$

which are further discussed in [17.2].

17.2.3 Quadrifilar Helix Antenna

A popular variant of a helical antenna, which is not restricted by the number of turns for circular polarization and, therefore, suitable for handheld GNSS terminals, is the quadrifilar helical antenna (QHA) [17.11]. It offers a simple geometry and small size, and is able to produce a circularly polarized hemispherical radiation pattern, but also exhibits some backlobes.

Unlike a traditional monofilar (single-wire) helix antenna, a QHA consists of four standalone helices fed with progressive quadrature phase shifts (Fig. 17.6). The rotation of circular polarization is defined by the curling direction of the monofilars while the increasing or decreasing phase progression dictates the direction of the radiation pattern (forward- or back-fire). The resonant QHA can have variable lengths, such as $1/4\lambda$, $1/2\lambda$, $3/4\lambda$, or a full wavelength λ [17.13]. Similarly, the number of turns can also vary from $1/4$ to 1. Moreover, for the volutes of an odd number of quarter wavelengths (i.e., $\lambda/4$ and $3\lambda/4$), the end of the helix element is open circuited, whilst for an even number of quarter wavelengths, the end of the helix element is bent

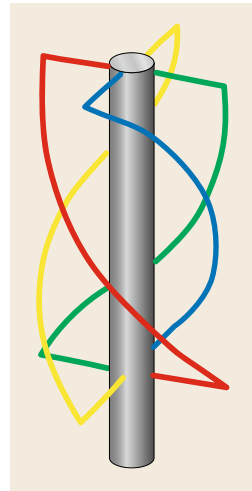


Fig. 17.6 Schematic view of a half-turn quadrifilar helix antenna. For improved clarity, different colors are used for the individual wires (after [17.12])

toward the center and short circuited to the ends of the other helix elements [17.13].

Despite the suitability of QHA for GNSS applications, its manufacturing is complex. A planar quadrifilar helical antenna (PQHA) simplifies the antenna fabrication while maintaining the performance of the traditional QHA. Etching the PQHA on the thin layer of flexible substrate material allows for line meandering, looping, bending, and varying the pitch angle, thereby reducing the size and mass [17.2].

Similar to microstrip antennas, the quadrifilar helix antenna can also be modified for multiband operation. A dual-band QHA can be constructed by running parallel lengths of helix-wound wires where the length of each wire determines the resonant frequency. Alternatively, a trap circuit can be inserted into each of the four wires to shorten their effective length at the higher signal frequency [17.14]. A folded QHA covering the full range of GNSS frequency bands as well as the adjacent IRIDIUM communication frequencies has been developed by the MITRE Corporation [17.15]. A tri-band PQHA has also been demonstrated by integrating a dual-band and a single-band PQHA in a piggy back configuration [17.16]. Although these antennas demonstrate good GNSS performance, there is yet a long way to go when these antennas will be used commercially.

One of the main issues with the use of the QHA is the relatively complicated phase requirements for the feeding network. In order to produce the circular polarization in the desired axial direction, a progressive phase shift between the multiple arms is required to be produced. This can be achieved in two different ways. Either a 1-to-4 feed network consisting of a 180° and 90° hybrids, as shown in Fig. 17.7, can be used

or there is also an alternative solution. The quadrifilar antenna is considered to be made of two bifilars combined together. By making one bifilar slightly longer than the second bifilar, a $\pm 45^\circ$ phase shift can be created, allowing two adjacent arms of the QHA to be connected to the same balun terminal, and still achieve the phase quadrature relationship necessary for correct operation [17.2].

17.2.4 Spiral Antenna

Spiral antennas are popular for their wide-band circular polarization operation. The antennas actually belong to a category called frequency-independent antennas. Spiral antennas come in more than one configurations such as Archimedean spiral, logarithmic (log) spiral, and conical spiral [17.1, 2]. The Archimedean and the log spiral are 2-D planar configurations and can be considered as a dipole antenna twirled in a spiral fashion. The conical spiral, on the other hand, takes its inspiration from a helix antenna that is tapered to form a conical shape.

Both planar and conical spirals can be manufactured easily using printed circuit board (PCB) techniques. The conducting conical spiral surface can be constructed by forming conical arms on the dielectric cone. The feed cable can be bound within the metal arm and wrapped around the cone. The radiation pattern is toward the apex with its maximum along the axis. Similarly, a planar spiral can also be etched on a thin substrate layer.

In this section, we will stick to the design of the log spiral only. Similar to a dipole, a planar spiral antenna has an omnidirectional but circularly polarized radiation pattern. Therefore, in order to achieve a semi-hemispherical pattern, a requirement for GNSS, a spiral antenna backed with a decent reflector is usually employed.

The general shape of a logarithmic spiral antenna can be modeled by the relation

$$r = R_0 e^{a\phi}, \quad (17.13)$$

which describes an exponential growth of the radius of the two spiral arms with polar angle ϕ in the antenna plane (Fig. 17.8). R_0 is the initial radius of the spiral and matches 1/4 of the shortest wavelength that can be transmitted or received by the antenna. On the other hand, the outer circumference ($2\pi R_{\max}$) matches the longest wavelength at which the antenna can be operated.

The factor a defines the rate at which the spiral grows with the increasing angle ϕ and effectively controls the spacing between the spiral arms. A small value of a will result in extremely small spacing between the spiral

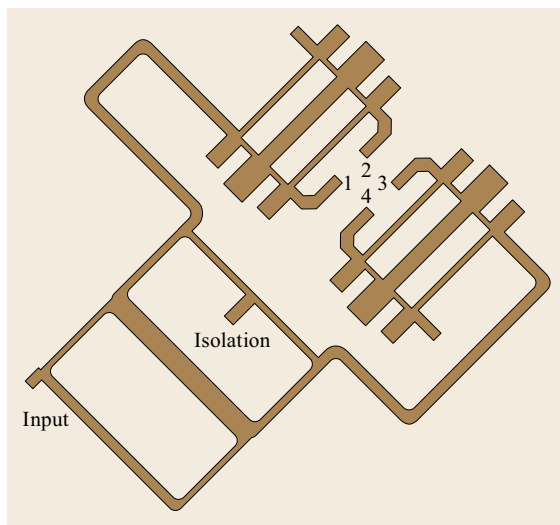


Fig. 17.7 1-to-4 hybrid coupler

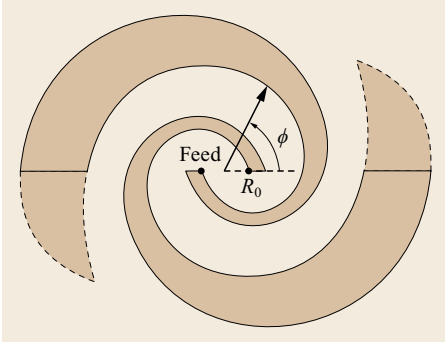


Fig. 17.8 Center-fed log spiral antenna (after [17.2])

arms, making the antenna behave like a capacitor with very poor radiation. On the other hand, a large value of a will make the spiral behave like a dipole losing its circular polarization capability. A commonly recommended value for a good spiral operation is $a = 0.22$ [17.2].

17.2.5 Wide-Band Bow-Tie Turnstile Antenna

Another type of antenna that can achieve a wide bandwidth is a bow-tie turnstile antenna. The antenna produces an omnidirectional radiation pattern and, therefore, requires a reflector ground plane to achieve a hemispherical radiation pattern. A bow-tie turnstile antenna, shown in Fig 17.9, has been used in [17.17]. The antenna is optimized to achieve 28% impedance bandwidth at the L-band.

It uses a differential feed and a 90° hybrid coupler to achieve circular polarization. Antenna measurement results show that the antenna achieves an average gain of 8 dBiC (i. e., 8 dB relative to a perfectly isotropic circularly polarized antenna) across the resonant bandwidth. The antenna has been tested by integrating it to a flat ground plane [17.17].

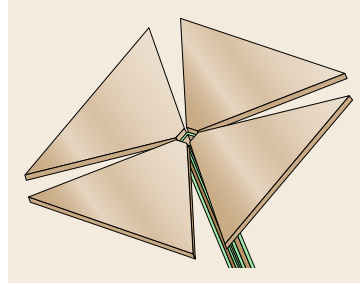


Fig. 17.9 Bow-tie turnstile antenna



Fig. 17.10 Modified NovAtel pinwheel antenna (after [17.18], courtesy of NovAtel)

17.2.6 Wide-Band Pinwheel Antenna

A wide-band pinwheel antenna designed by NovAtel, Inc. has been presented in [17.18, 19]. The antenna configuration consists of an array of coupled spiral slots arranged in a pinwheel of about 20 cm diameter backed by a microstrip-based multiple-turn spiral transmission line (Fig. 17.10). The overall antenna structure is compact, lightweight, and has multipath mitigation capability for almost the entire L-band. However, antenna measurement results presented in [17.18] show that the antenna performance is not uniform across the resonance bandwidth with better performance achieved at higher frequency. The antenna achieves a 6.8 dBiC gain at 1.575 GHz (L1) but only 2.8 dBiC at 1.227 GHz (L2).

17.3 Application-Specific GNSS Antennas

The overview of antenna concepts given in the previous section illustrates that any antenna type with a circularly polarized semihemispherical pattern can, in principle, be used for GNSS. However, it should be kept in mind that not all antennas are equally suitable for all types of GNSS applications [17.20]. As an example, antennas for handheld terminals will require a very low form factor, while superior (PC) stability is required for surveying and geodetic applications. Even more specialized requirements may apply for scientific applications such as GNSS remote sens-

ing and atmospheric sounding. Table 17.2 presents an overview of common GNSS applications and related antenna requirements. Specific antenna solutions for some of these fields are discussed throughout this section.

17.3.1 Hand-Held Terminals

Hand-held terminals (such as mobile phones) require a high level of miniaturization. Two specific examples of microstrip patch and helix antennas are presented in

Table 17.2 Application-specific GNSS antenna requirements (after [17.21])

	Low profile	Low PCO/PCV	Rugged	Multi-constellation	High Multipath suppression	Pole mount	Magnetic/surface mount	Extended temperature range	High altitude operation
GIS		✓	✓	✓		✓			
Survey		✓	✓	✓	✓	✓			
Reference station		✓	✓	✓	✓	✓		✓	✓
Vehicle tracking	✓						✓		
Construction/mining	✓	✓	✓	✓	✓	✓	✓		
Precision agriculture		✓		✓			✓		
Marine				✓	✓	✓			
Aviation	✓	✓		✓				✓	✓
Unmanned aircraft	✓	✓		✓				✓	✓
Timing						✓			

GIS – geographic information system

the following sections. Even though both antenna concepts can offer a very small form factor, the microstrip antenna is often preferred since it can be etched using simple PCB technology and enables extremely thin designs. However, as the antenna needs to pick up all signals from above the horizon, the radiation pattern and the direction of the main lobe are likewise important criteria for the antenna selection.

Inverted-F Antenna

An antenna type very commonly used for GPS in mobile phones is an inverted-F antenna (IFA). The antenna produces an omnidirectional pattern covering all visible GNSS satellites. The antenna is linearly polarized and, therefore, able to receive both RHCP (direct line of sight) and LHCP (reflected) GNSS signals. This may be considered as a problem in the conventional GNSS as the reflected signals tend to degrade position accuracy. However, handheld terminals are normally operated near the ground, where the path difference between direct and reflected signals results in only a limited performance degradation. Moreover, most handheld terminals only require a moderate level of accuracy and can therefore tolerate such performance limitations. Figure 17.11 shows the inverted-F GNSS antenna for a Samsung Galaxy tablet.

Dielectric Loaded Quadrifilar Helical Antenna

A miniaturized GNSS antenna for handheld terminals presented in [17.22] is shown in Fig. 17.12. The antenna is a compact version of a PQHA, where the small size is achieved by wrapping the QHA around a high-permittivity material. The antenna gives a uniform semihemispherical pattern above the ground and uses a sleeve balun in order to balance the unbalanced current in the coaxial line. Since the antenna is tightly wrapped around a high-permittivity material, its gain is

very less and therefore a low noise amplifier (LNA) is directly integrated to the antenna feed.

17.3.2 Surveying and Geodesy

Surveying (Chap. 35) and geodesy (Chap. 36) represent two high-end GNSS applications, which make use of advanced carrier-phase-based techniques to achieve mm-level relative or absolute positioning accuracies. In order to fully exploit the precision offered by GNSS carrier-phase measurements, adequate care must be taken in the choice and design of the antenna. Key requirements for geodetic-grade GNSS antennas include a high level of multipath rejection as well as phase center stability. Since dual-frequency observa-

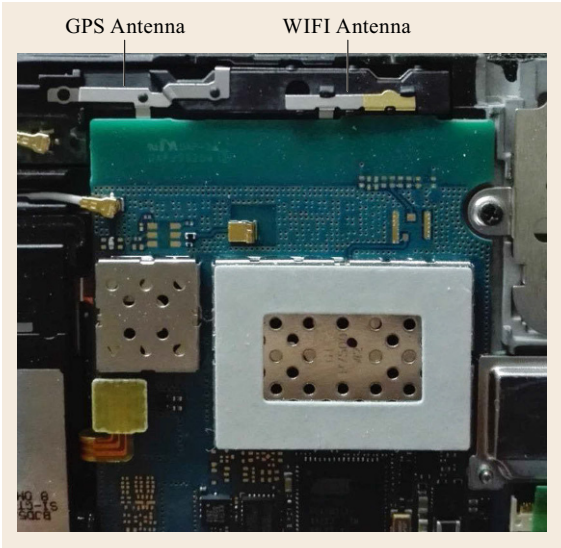


Fig. 17.11 Inverted-F GPS antenna as used in a Samsung Galaxy tablet computer (courtesy of M. Maqsood)



Fig. 17.12 GNSS antenna for handheld terminals. The printed quadrifilar helix (PQFA) has a diameter of about 10 mm and is integrated with a low-noise amplifier (LNA)

tions are required to eliminate ionospheric path delays in precise point positioning or long-baseline differential GNSS techniques, the employed antennas must maintain its high performance in multiple individual frequency bands or even the full range of GNSS frequencies.

Multipath errors arise whenever the direct signal received from a GNSS satellite is superimposed by reflected signals and affects the code and carrier tracking inside the receiver. As discussed in Chap. 15, multipath effects can partly be mitigated through advanced correlators and tracking techniques. However, a careful antenna design may help to avoid reflected signals from entering the antenna in the first place and thus offers one of the most efficient means of protection against multipath-induced GNSS measurement errors.

The use of a *choke ring* (Sect. 17.4) around the sensitive antenna element constitutes one of the most widely employed techniques to reduce the adverse impact of low-elevation, reflected signals, even though it is less convenient for nonstationary applications. An example of such a choke-ring antenna for geodetic reference stations is shown in Fig. 17.13. To protect against rain or snowfall, the antenna is typically covered by a radome as illustrated in the right part of the image. In view of their significance, choke rings and other concepts of multipath-mitigating antennas are further discussed in a subsequent section (Sect. 17.4) of this chapter,

Next to proper multipath reduction, a high stability of the antenna constitutes the second main criterion for antennas used in high-precision positioning applications. Common specifications for surveying antennas require PCVs of less than a few millimeter across the relevant range of elevations in each of the supported frequency bands. Even though it has become common



Fig. 17.13a,b Choke-ring antenna (Leica AR25) without (a) and with (b) protective radome (courtesy of O. Montenbruck/DLR)

practice to calibrate PCVs of geodetic-grade antennas for utmost positioning accuracy (see Fig. 17.14 and Sect. 17.6.2), the use of such calibration tables may not be convenient or feasible in all applications. Low PCV amplitude and good PCV repeatability among different units of a given antenna model therefore remain important antenna criteria for high-end GNSS positioning. For completeness, we note that these criteria also apply for the radome used to protect the actual antenna from environmental effects. Even though the radome is from nonconductive material such as plastic, it impacts the radiation pattern and care must be taken to minimize (and/or properly calibrate) associated PCOs shifts and PCVs.

17.3.3 Aviation

GNSS-based navigation for general aviation is so far limited to the use of the GPS SPS along with satellite- or ground-based augmentation systems (SBAS/GBAS) to enhance the integrity of the position information (Chap. 30). As such, aviation antennas are commonly confined to the L1 frequency, even though various types of dual-frequency antennas are available for military users. However, a growing need for dual-frequency aviation antennas will emerge along with the ongoing introduction of L5/E5 GPS, Galileo, and SBAS signals for safety critical aviation applications.

Compared to personal navigation antennas, miniaturization is not required for aviation use, but environmental robustness becomes a driving factor for the antenna design. In particular, the antenna should be able to operate at extreme vibration conditions and should be rugged enough to sustain a wide range of temperatures. Furthermore, the accommodation on the aircraft calls for a low profile or even conformal design to minimize drag and turbulences. Finally, aviation antennas must be tolerant to lightning to enable safe operation under severe weather conditions. Relevant standards for environmental robustness (DO-160 [17.24]) and mini-

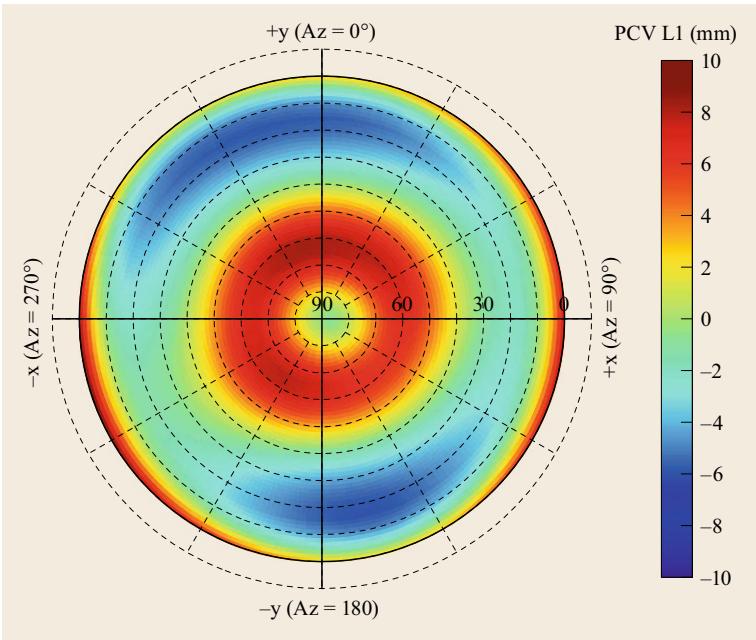


Fig. 17.14 Phase-center variations of a geodetic-grade choke-ring antenna (Trimble TRM14532.00) on the L1 frequency as derived from robot calibrations (after [17.23], courtesy of igs08.atx antenna model)

mum performance (DO-208 [17.25]) of airborne GNSS equipment have, for example, been issued by the Radio Technical Commission for Aeronautics (RTCA). An example of an aviation GPS antenna with a standardized form factor and mounting holes is shown in Fig. 17.15.

17.3.4 Space Applications

GNSS receivers are nowadays widely used on spacecraft in the low Earth orbit as a means for onboard navigation and timing as well as precise orbit determination (Chap. 32). Similar to terrestrial antennas, the choice of an antenna for use onboard a spacecraft is largely dependent on application-specific needs

and constraints such as the desired measurement quality and the available space for antenna accommodation.

Standard patch antennas encapsulated into an appropriate radome (Fig. 17.16a) are widely used for code-based single-frequency navigation that do not require the utmost precision of the resulting measurements. Even though many of these antennas have been specifically designed for use in space use, aviation antennas have also been found to be viable and cost-effective alternatives for many missions.

For precise orbit determination, dual-frequency antennas with well-calibrated phase centers and low multipath sensitivity are required. This is commonly accomplished by combining a patch or a cross-dipole antenna element with a choke ring. This choke ring is similar in function to those in geodetic reference sta-



Fig. 17.15 L1 GPS/SBAS antenna for general aviation (courtesy of Sensor Systems, Inc.)

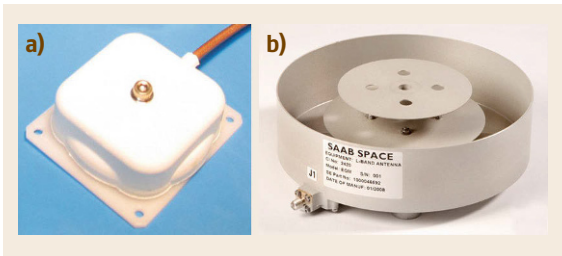


Fig. 17.16a,b Examples of GPS antennas for space use: (a) L1 patch antenna with radome (courtesy of SSTL), (b) dual-frequency patch excited cup antenna (courtesy of RUAG Space AB). Images not to scale

tions, but requires special effort in the manufacturing to meet the tight mass budgets and qualification standards of a space mission. Examples of such antennas as used on the Gravity Recovery And Climate Experiment (GRACE) and TerraSAR-X missions are given in [17.26] along with a discussion of ground and in-flight calibrations of their PCs.

An alternative dual-band, multipath mitigating antenna design has been proposed in [17.27]. The patch excited cup (PEC) antenna uses a stacked patch configuration mounted in a cup-shaped ground plane. It enables dual-band operation, while keeping the spacecraft interference to the antenna at a minimum. An example of a PEC antenna flown on the SWARM satellites is shown in Fig. 17.16b. A further enhancement of the multipath suppression capability can be achieved by adding a small number of choke rings inside the cup [17.28] as used for the Sentinel mission. Compared to traditional choke-ring antennas, the PEC design offers a high performance at a notably lower form factor.

Apart from its RF characteristics, a GNSS antenna for spaceborne applications needs to account for the specific environment in which it is operated [17.29]. This includes extreme temperatures, vacuum, and possibly space radiation. All materials for manufacturing the antenna should be space qualified and issues such as low outgassing need to be considered in the material selection. Where possible preference, should be given to materials that have previously been flown and tested in space. The antenna structure must be robust enough and the mounting of the antenna onto the satellite body has to be very rigid to survive shocks and vibration during the launch. Furthermore, the presence of other subsystems around the antenna and the mutual coupling between the antenna and the spacecraft body have to be taken into account. It is also important to avoid any suspended metallic layers within the antenna structure since charge deposition on the metallic structure of the antenna may cause harm to the LNA and other circuits.

17.3.5 Antijamming Antennas

Controlled reception pattern antennas (CRPA) represent a special type of beam-forming antenna, in which the shape of the beam pattern can be actively controlled and adapted during the operation [17.30]. While originally conceived for military users as a means to mitigate intentional jamming or spoofing of GNSS signals (Chap. 16), the use of CRPAs is getting of wider interest for civil users in the context of assuring position, navigation, and timing (PNT) services for critical infrastructure.

While fixed radiation pattern antennas (FRPAs) as used in various GNSS remote sensing applications mainly aim to maximize the antenna gain pattern in a given region of interest (e.g., toward the limb of the Earth; see Sect. 17.3.6), antijamming antennas adaptively steer selected pattern nulls in the directions of interference sources or adverse signal reflections. This helps the GNSS receiver to maintain the required signal-to-noise ratio, since the contribution of the interference signal is attenuated without affecting reception of the desired GNSS signals.

To achieve this goal, the signals of multiple antennas need to be phase coherently combined. This is illustrated in Fig. 17.17 for a simple set of two antennas with a PC separation of $\lambda/2$ and two delay elements that can be adjusted within the limits

$$0 \leq \delta\phi_1 - \delta\phi_2 \leq +\frac{\pi}{2}.$$

Using complex notation, the combined signal can be described as

$$\begin{aligned} S &= A \left(e^{2\pi j \left(\frac{\cos(E)}{2} + \delta\phi_1 \right)} + e^{2\pi j (\delta\phi_2)} \right) \\ &= A e^{2\pi j \delta\phi_2} \left(1 + e^{2\pi j \left(\frac{\cos(E)}{2} + \delta\phi_1 - \delta\phi_2 \right)} \right), \end{aligned} \quad (17.14)$$

where A denotes the single-antenna signal amplitude and E is the elevation angle. For

$$\delta\phi_1 - \delta\phi_2 = \frac{1}{2}(1 - \cos(E))$$

the total phase shift between both signals prior to combination amounts to half a wavelength, which results in a complete cancellation of signals received at this el-

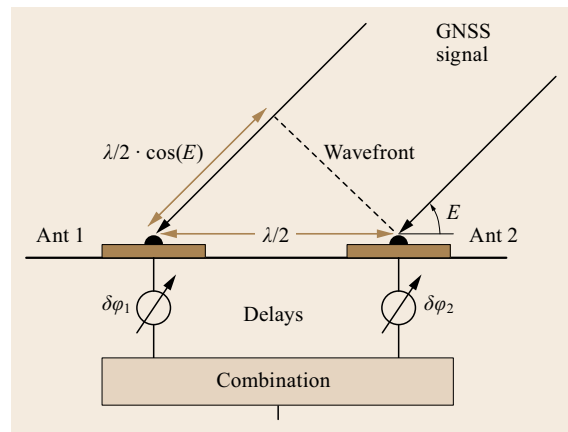


Fig. 17.17 Schematic view of GNSS signal cancellation using two combined antennas

evaluation. For the given antenna geometry, exactly one null in the antenna diagram can be generated at a configurable direction in the plane of the illustration. This basic concept is easily extended to three dimensions using an array of 2×2 antenna elements, which then allows steering of the gain minimum throughout the entire hemisphere.

The common design strategy of the CRPAs is to put the individual antennas in the form of a circular array creating the nulls to be pointed in the direction of potential jammers. An important consideration in the design of a CRPA is the number of antenna elements as the number of nulls generated is usually one less than the number of individual elements [17.31]. However, this may come at the cost of reduced satellite visibility as increasing the number of antenna elements to introduce nulls reduces the antenna beam width [17.32]. Moreover, the large antenna array and complex electronics package increase the overall antenna mass and size. Compact arrays as described, for example, in [17.33, 34] offer better portability but are more challenging in terms of the required isolation between individual elements. An example of a seven-element antenna array for beamsteering and interference suppression is shown in Fig. 17.18.

Even though it is possible to build a CRPA as standalone unit that can be connected to arbitrary receivers like any other antenna, jamming-resistant GNSS terminals often involve a closely coupled design of the antenna elements and the receiver electronics. This is due to the fact that a phase-coherent combination of signals with variable delay from multiple input antenna elements can best be accomplished within the actual signal-processing chain. The use of software-defined radio techniques (SDR) and the ability to implement flexible signal-processing concepts in field programmable gate arrays (FPGAs) has enabled great

progress in the design of jamming-tolerant GNSS receivers with multiantenna arrays for single- and dual-frequency applications.

The detailed design of CRPAs for GNSS anti-jamming applications and relevant beamforming algorithms exceed the scope of this chapter, but are readily covered in publications such as [17.30, 31, 35–37] along with practical examples and test results.

17.3.6 GNSS Remote Sensing

GNSS remote sensing [17.38] refers to a category of applications, where signals coming from the GNSS satellites are used to study diverse properties of the Earth's surface and atmosphere. While some of these applications do not require specific equipment, GNSS reflectometry, scatterometry, and radio occultation measurements are usually performed with specialized GNSS receivers as well as dedicated antennas.

GNSS reflectometry and scatterometry (Chap. 40) make use of GNSS signals reflected or scattered from oceans, land, or ice to retrieve information, such as wind speed, salinity, soil moisture, and ice-layer density, or even contribute to the detection of tsunami waves for disaster monitoring. The techniques resemble traditional radar and altimeter observations, but make the use of GNSS signals of opportunity and do not need an active illuminator as part of the instrumentation.

In the context of antenna design, a GNSS reflectometry antenna is usually a high-gain array antenna with an off-nadir pointing beam. A high-gain beam is required as the antenna needs to pick up weak reflected signals from a high-altitude (air- or spaceborne) platform, while the off-nadir pointing ensures a larger observation area. Since the polarization of the GNSS signals changes upon reflection, the reflectometry antenna is designed for LHCP signal reception.



Fig. 17.18 Seven-element beamsteering antenna for multi-band GPS, Galileo, and GLONASS reception (courtesy of DLR Institute of Communications and Navigation)

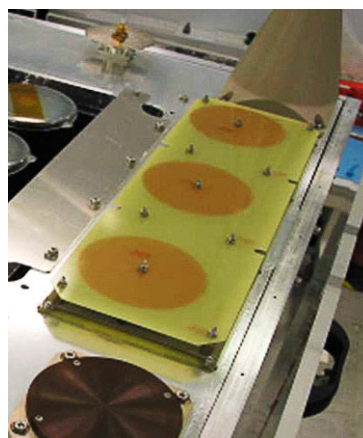


Fig. 17.19 GNSS reflectometry antenna array of the UK-DMC satellite (courtesy of SSTL)

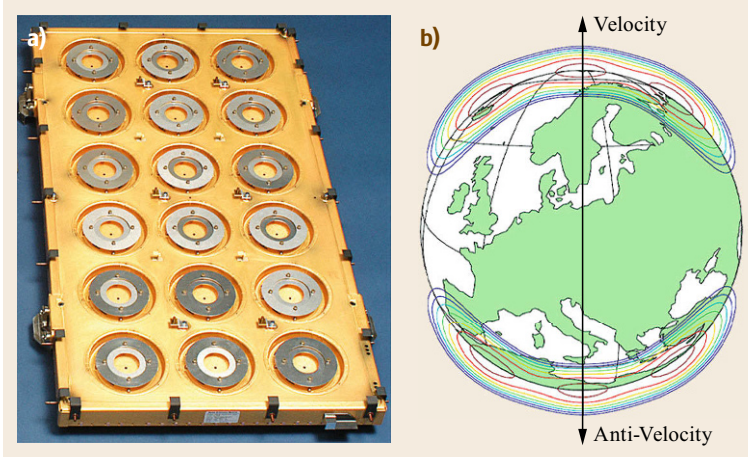


Fig. 17.20a,b Beam-forming array antenna for radio occultation measurements onboard the Metop satellites. **(a)** Separate antennas are used for the forward (velocity) and aftlooking (antivelocity) direction. The gain pattern of each antenna is adjusted to cover a narrow region close to the Earth's limb, where occultations from rising or setting GPS satellites can be observed **(b)** (courtesy of RUAG Space AB)

Figure 17.19 shows a three-element microstrip antenna array used as part of an early scatterometry experiment onboard the UK-DMC satellite [17.39]. The antenna works at the L1-band and follows a stacked configuration to achieve a greater than 12 dBiC gain. The design of a dual-frequency antenna array for spaceborne reflectometry applications is described in [17.40]. It has first been applied in combination with SSTL's SGR-ReSI (space GNSS receiver remote-sensing instrument) onboard the UK TechDemoSat-1 satellite [17.41].

Next to reflectometry, radio occultation measurements represent another GNSS remote sensing technique, which depends on specialized antenna designs for optimum performance. Radio occultations make use of signals propagating through the deep layers of the atmosphere between the transmitting GNSS satellite and a receiving satellite in low Earth orbit (Chap. 38). The GPS/Meteorology (GPS/MET) mission [17.42] has first shown that radio occultation can be used to predict accurate, all weather, global refractive index, pressure, density profiles in the troposphere, as well as electron density profiles and total electron content (TEC) of the ionosphere. Today, radio occultation measurements are routinely conducted by a variety of Earth orbiting satellites, which contribute directly to near-real-time weather prediction services.

Tropospheric radio occultation measurements are most challenging due to the pronounced attenuation experienced by the signals through the dense layers of the atmosphere. The situation is further complicated by severe losses experienced in the semi-codeless tracking of the GPS P(Y)-code signals that are still most widely used today, since a civil L2 signal is not yet available on the entire GPS constellation. Similar to reflectometry, the atmospheric sounding, therefore, benefits from array antennas, which provide a high gain in the direction of the horizon.

Figure 17.20 shows an 18-element array antenna for use with the global navigation satellite system receiver for atmospheric sounding (GRAS) onboard the metop satellites [17.43]. Each antenna element consists of RHCP concentric rings for dual-frequency (L1/L2) signal reception. The individual elements consist of a pair of concentric rings and are mounted in a cup in order to reduce mutual coupling and mitigate multipath interference. The ring elements are slot-fed, and an optimal performance for both frequency bands is obtained by using separate feed networks for each band [17.44]. Compared to smaller arrays used on other radio occultation missions, the large antenna array and sharp beamforming of the GRAS antenna contributes to a superior measurement quality achieved by the Metop satellites.

17.4 Multipath Mitigation

Multipath interference is one of the major error sources in GNSS, as it can severely degrade the accuracy of GNSS measurements. It arises when multiple copies of the same signal arrive at the receiver from more than one directions. This happens when the primary line-of-

sight (LOS) signal suffers reflection or diffraction from the antenna surface or nearby objects. These attenuated and phase-shifted (delayed) signals are summed with the LOS signal in the receiver, which results in distortions of the correlation function and associated

errors in the range and carrier- phase measurements (Chap. 15).

A common multipath scenario is illustrated in Fig. 17.21 for a pole-mounted antenna. Here, the excess path difference

$$d = 2h \sin \theta, \quad (17.15)$$

between the direct signals and the signal reflected from the ground is usually much smaller than the GNSS code chip length. The receiver correlator may thus be unable to discriminate between the direct and reflected signals.

Multipath interference can be mitigated to a large extent by shaping the radiation pattern to have a sharp gain roll-off toward the horizon and by suppressing the sensitivity to signals received from the back of the antenna [17.45]. This helps in rejecting the reflected and diffracted signals coming from low- or even negative-elevation angles. Integration of external ground planes such as choke rings is amongst the popular methods for multipath mitigation at the antenna [17.28, 46, 47]. Alternatively, an antenna can be specifically designed to suppress the surface wave propagation [17.48–50], a stimulus for multipath interference, within the antenna structure. The present section discusses some of these techniques that are used to mitigate multipath in high-performance antennas. An overview of different multipath mitigation methods for GNSS antennas and an assessment of their performance are presented in [17.51].

17.4.1 Metallic Reflector Ground Plane

The simplest way to reduce the backlobe of an antenna is to put a metallic reflector underneath it. The metallic reflector will shield the antenna from the signals coming from below the horizon. A microstrip antenna can be directly mounted to a large reflector ground plane for performance improvement but generally most of the

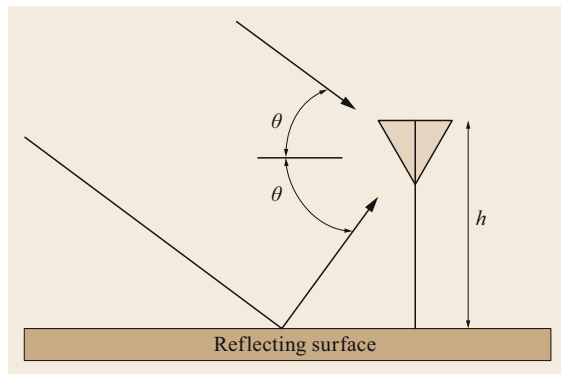


Fig. 17.21 Typical multipath scenario for a ground-based GNSS antenna

dual-frequency microstrip antenna designs have an integrated feed network at the most bottom. This makes the distance between the antenna base and the metallic reflector to be a critical parameter as putting it very close to the feed would generate higher order modes transforming the microstrip feed into a stripline configuration. Normally, the reflector is placed quarter of a wavelength ($\lambda/4$) away from the antenna base. Any electromagnetic energy coming toward the reflector would suffer a phase rotation of 180° , thereby forcing the antenna backlobes to be at minimum level.

17.4.2 Choke-Ring Ground Plane

Corrugated surfaces are widely used as ground planes (chokerings) to enhance the antenna performance. Their design is similar to corrugated horns, where metal corrugations are used to reduce the spill-over beyond a desired flare angle. Choke-ring ground planes help to sharpen the gain slope of the antenna radiation pattern, decrease side- and backlobe levels, thereby improving cross-polarization performance, and hence stabilize the antenna phase center. This is achieved by the cancelation of incoming waves after passing through successive corrugations as shown in Fig. 17.22.

Another interesting property of the corrugated surface is its high-surface impedance, which blocks the propagation of surface waves within the antenna substrate. Surface waves can occur on the interface between two dissimilar materials such as metal and air and travel along the boundary of the two surfaces. The high impedance of the choke ring is due to the quarter-wavelength deep corrugation which transforms the short-circuit boundary condition at the bottom of the corrugation to an open circuit at the top of the surface. Choke-ring ground planes are widely used with microstrip patch or Dorne–Margolin vertical dipole antenna elements for geodetic reference antennas with high PC stability and low multipath sensitivity.

The depth of individual grooves needs to be larger than one-quarter and less than one-half of the signal

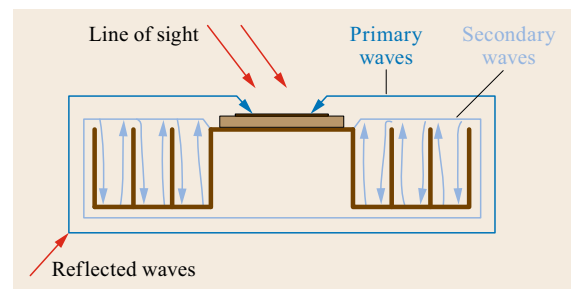


Fig. 17.22 Field waves in a choke ring (after [17.46])

wavelength [17.17] for a *cut-off* choke ring that fully suppresses the propagation of surface waves. Within this range, the lower value is preferred because of an optimum performance and smaller size of the resulting choke ring. To cover both the GPS L1 and L2 frequencies, traditional choke rings exhibit grooves of about $\lambda_{L2} = 61$ mm and are thus best suited for the L2 signal. The implementation of a dedicated dual-frequency choke ring design that achieves different groove depths for L1 and L2 through the use of a frequency-dependent diaphragm is described in [17.46].

17.4.3 Noncutoff Corrugated Ground Plane

A corrugated ground plane with corrugation depth smaller than $\lambda/4$ is presented in [17.17] to achieve multipath mitigation at low-elevation angles. The ground plane, as shown in Fig. 17.23, achieves the surface wave suppression in a different manner than traditional choke-ring. Instead of creating a high-impedance region around the antenna, the shallow corrugations allow the surface waves to propagate toward the ground plane boundary, where they become out of phase with the line-of-sight signals and get canceled.

Although the simulation and measurement results presented in [17.17] show that the shallow corrugations can achieve respectable multipath mitigating performance, the ground plane requires a higher number of successive corrugations. This results in an increased diameter which partly compensates the benefit of the decreased choke-ring height.

17.4.4 Convex Impedance Ground Plane

The integration of the choke-ring ground plane with an antenna creates an impedance surface parallel to the antenna ground plane. Although this improves the multipath mitigation capability of the antenna, it also reduces the antenna gain at the grazing angles, thereby

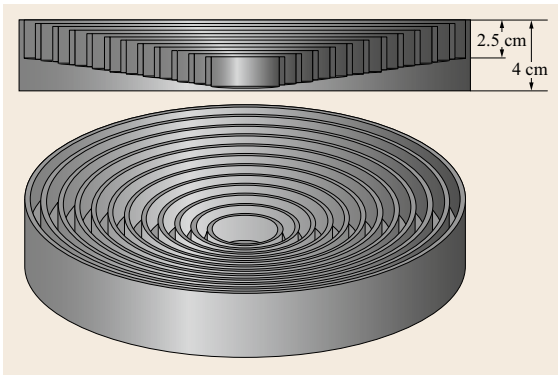


Fig. 17.23 Noncutoff corrugated ground plane

restricting the visibility of the low-elevation satellites. This happens as the grazing angles of an antenna with a flat ground plane coincide with the horizon. However, converting the flat ground plane to a convex one improves the antenna radiation pattern near the horizon while keeping the multipath mitigating capability.

A convex ground plane with pin structure has been presented in [17.52, 53] to achieve multipath capability over the GNSS band. The antenna design is shown in Fig. 17.24 where it uses an array of vertical convex dipoles capacitively coupled to the antenna feed. The overall antenna shows improved performance with regards to the visibility of low-elevation satellites.

17.4.5 3-D Choke-Ring Ground Plane

Despite the robust design and unique performance of the traditional choke-ring ground plane, it covers only a limited frequency range and provides poor sensitivity for satellites near the horizon. To cope with this limitation, a 3-D choke-ring ground plane has been proposed in [17.54]. Different from a conventional choke-ring ground plane, the 3-D choke ring has rings at multiple levels shaping it like a pyramid. The slots in the concentric rings allow the dissipation of unwanted RF energy. Similar to the convex impedance ground plane, a curved ground plane profile improves gain at grazing angles while keeping the PC stable.

The NovAtel GNSS-750 multi-GNSS reference station antenna shown in Fig. 17.25 integrates the 3-D choke-ring antenna with an ultra wide-band Dorne-Margolin antenna element. Test results for the build-identical Leica AR25 antenna [17.55, 56] confirm a notably improved low-elevation tracking along with a high PC stability over the entire frequency range, even though the forward-backward ratio is somewhat lower than that for a conventional 2-D choke-ring antennas.



Fig. 17.24 A convex impedance ground plane (courtesy of Topcon)



Fig. 17.25 Three-dimensional choke ring (courtesy of NovAtel)

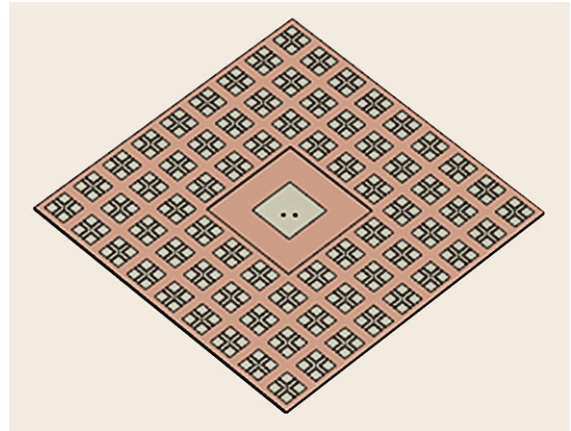


Fig. 17.27 Dual-band patch integrated with EBG structure

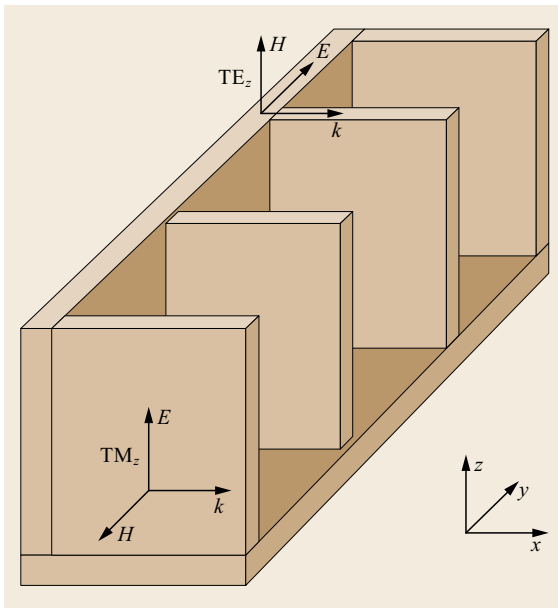


Fig. 17.26 Vector representation of TE_z and TM_z waves on the cross-reflector ground plane surface (after [17.47])

17.4.6 Cross Plate Reflector Ground Plane

The design of a cross plate reflector ground plane (CPRGP) for multipath mitigation is discussed in [17.57]. Here, the antenna element is surrounded by a square of crossed sidewalls with a shape shown in Fig. 17.26.

Other than a flat perfect electric conductor, which only cancels the transversal electric (TE_z) waves, the

cross plate reflector also cancels the transversal magnetic (TM_z) waves. This enables the reduction of the overall ground plane thickness to less than $\lambda/4$, that is, almost half that of a traditional choke ring. A prototype implementation of a dual-band CPRGP antenna using a shorted annular ring antenna element is presented in [17.47]. Tests confirm favorable multipath characteristics capabilities at a compact size of $19 \times 19 \times 3 \text{ cm}^3$.

17.4.7 Electromagnetic Band Gap (EBG) Substrate

Electromagnetic band gap (EBG) substrate [17.58] is a fairly new technique for surface wave suppression and has recently been used in GNSS for multipath interference mitigation. EBG surfaces are made by periodic structures producing very high surface impedance within a specified frequency band called the band gap where the high-impedance property is usually exploited for surface or lateral wave suppression [17.59].

An EBG substrate can be characterized by two properties, namely the employed material and the design of the EBG cells. A common EBG substrate design known as mushroom cell has properties similar to that of a corrugated metal surface. It blocks the propagation of surface waves by introducing a high-impedance region around the antenna. The high impedance is achieved by the two-dimensional distributed resonant circuit where capacitance is produced by the proximity coupling of periodic patch elements (mushroom hats in the mushroom unit cell) and the desired inductance is achieved by making loops of currents using via stems. A GNSS patch antenna integrated within an EBG surface is presented in Fig. 17.27.

17.5 Antennas for GNSS Satellites

Other than common GNSS receive antennas, which aim at a largely hemispherical gain pattern, the transmit antenna onboard a GNSS satellite needs to be highly directive to illuminate the Earth from a high altitude. While this could, in principal, be achieved with a narrow beam helix antenna pointed toward the nadir direction, such a solution poses two major disadvantages. Firstly, a notable fraction of the radiated energy would still pass by the Earth, if the gain variation across the visible surface of the Earth is kept small. Secondly, users in the nadir direction would always receive a higher flux than users near the rim of the Earth, since they benefit from both a smaller distance and a higher antenna gain.

Considering, as an example, the global positioning system with an orbital radius of $a = 26\,560$ km, the distance to the receiver varies between $r_{\min} = 20\,180$ km for a user (A) that sees the GPS satellite at zenith and $r_{\max} = 25\,780$ km for a user (B) that sees the GPS satellite near the horizon (Fig. 17.28). This corresponds to a factor of 1.6 or, equivalently, a 2.1 dB difference in the associated free-space loss.

17.5.1 Concentric Helix Antenna Arrays

To cope with this type of problem, shaped antenna patterns obtained by the combination of multiple antenna elements arranged in concentric rings have earlier been proposed for geostationary communication satellites [17.60]. They are based on the consideration that an M-shaped far-field antenna pattern with a local minimum along the boresight direction and a maximum near the apparent semidiameter of the Earth can be achieved by a specific choice of the illumination function across the antenna aperture. More specifically, the radial variation of the illumination function should be described by $J_1(x)/x$, where J_1 is the first-order Bessel function

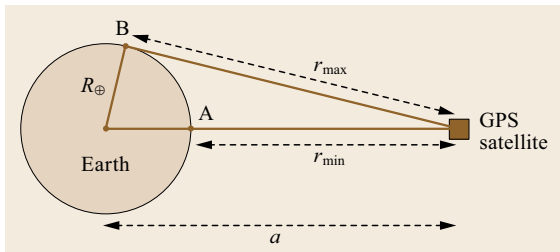


Fig. 17.28 Geometric conditions for a GPS satellite transmitting navigation signals to the entire surface of the Earth. Points on the surface of the Earth with the minimum and maximum distance from the transmitting satellite are marked as A and B, respectively

and $x = d/\lambda$ is the radial distance from the center of the aperture in units of the wavelength.

As discussed in [17.62], the ideal illumination function can be approximated by two concentric rings with suitably chosen radii that is populated with a total of 12 individual helix antenna elements (Fig. 17.29). The array comprises an inner ring of four antennas, which contributes the central peak of $J_1(x)/x$, while the outer antenna ring is placed at the first local minimum near $x = 4.6$. By feeding the outer antenna ring with a relative phase shift of 180° and using a reduced power for each antenna compared to the inner ring, a dip along the boresight direction is ultimately achieved in the far-field radiation pattern.

This concept was first implemented on the GPS Block I satellites [17.61] and successfully applied since then on all subsequent generations. It is also widely employed on other GNSS satellites as illustrated in Fig. 17.30 for selected GPS and Quasi-Zenith Satellite System (QZSS) satellites. The L-band antennas on the GPS Block IIR satellites comprise slim helices similar to the earlier Block I and II/IIA satellites, but are interleaved with an array of (shorter and thicker) ultra-high frequency (UHF) helix antennas on these satellites. A 4-plus-8 helix antenna array is also used by the GLONASS satellites. In the case of the latest K1 series, as shown in the figure, the free space between the inner and outer ringd is populated with retro-reflectors for satellite laser ranging. Finally, the figure shows the main L-band antenna of the QZS-1 satellite. It covers

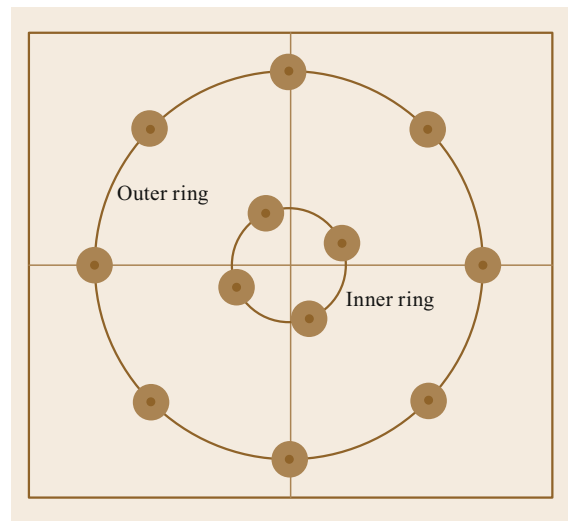


Fig. 17.29 Arrangement of the 12 antenna elements on the GPS transmit antenna panel. The two rings have diameters of about 30 and 90 cm (after [17.61])

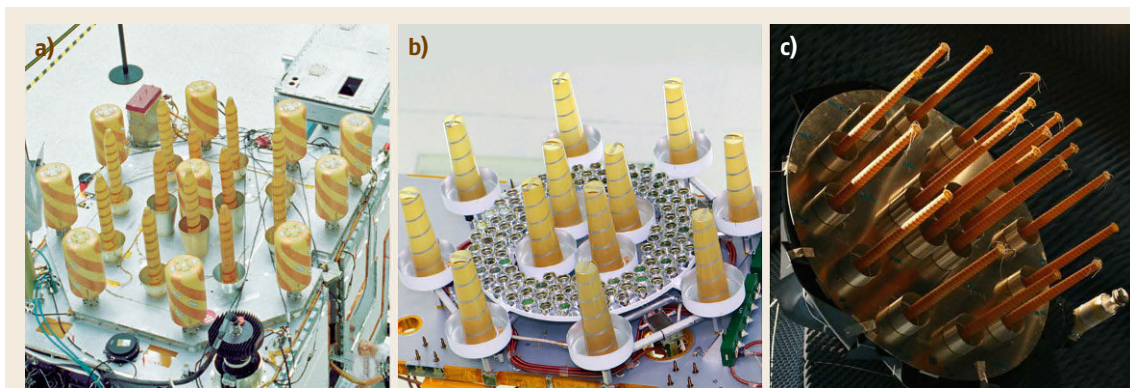


Fig. 17.30a–c Helix antenna arrays as used on the GPS Block IIR satellites (a), the GLONASS-K1 satellites (b) and the QZS-1 satellite (c) (courtesy of Lockheed Martin (a), ISS Reshetnev (b), and (JAXA) (c))

a total of four frequencies (L1, E6, L2, and L5) and is made up of a larger number of individual helix elements arranged in a central disk and an outer ring.

Measured antenna patterns are shown in Fig. 17.31 for two current types of GPS satellites (Block IIR and IIR-M) with slightly different antenna panels. A boresight angle of 0° refers to the nadir direction, while boresight angles of 14° mark the limb of the Earth as seen from the GPS satellite. The characteristic M-shape is most obvious for the L1 gain pattern of the Block IIR satellites, while an almost flat pattern over the entire surface of the Earth is obtained at the L2 frequency. Compared to the legacy IIR panel, the modernized panel of the IIR-M satellites uses new helix elements and a slightly modified arrangement, which results in the slightly modified overall gain pattern.

A diagram of the IIR-M gain pattern covering the entire hemisphere is shown in Fig. 17.32. Aside from the central main lobe, various sidelobes can be recognized at boresight angles near 30° and 55° at four distinct azimuth angles that reflect the four-quadrant symmetry of the antenna panel. The sidelobes point far off the Earth and are thus of no relevance for terrestrial and aeronautical navigation. However, their presence notably increases the availability of navigation signals for spaceborne users in high-altitude orbits around the Earth [17.63].

17.5.2 Patch Antenna Arrays

As an alternative to the helix antenna arrays discussed above, the European Galileo system makes use of flat antenna arrays with a large number of individual patch elements to obtain the desired isoflux phase pattern over the E1, E6, and E5 frequency bands. Two different base designs are used on the individual satellites of the Galileo family. They include a hexagonal

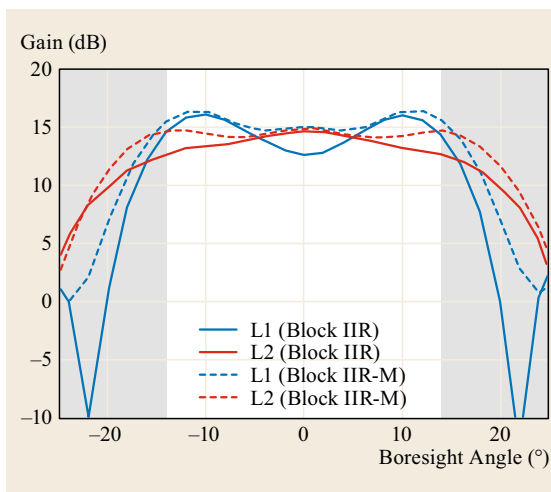


Fig. 17.31 Gain pattern of GPS Block IIR and IIR-M satellites on the L1 and L2 frequencies (after [17.64]). Shaded areas mark regions outside the apparent Earth disk

array of circular photo-printed patch elements developed by EADS CASA and used on Galileo in-orbit validation element (GIOVE)-B as well as the Galileo in-orbit validation (IOV) satellites as well as a near-circular array of four-layer stacked patches developed by Thales Alenia Space and used on the GIOVE-A and Galileo full operational capability (FOC) satellites (Fig. 17.33).

The Galileo IOV antenna consists of a densely populated array of photo-printed stacked patches generating a dual-band circular polarization. While 42 antenna elements were used on the GIOVE-B prototype, the number was later increased to 45 for the IOV satellites. The hexagonal array is formed by combining six sectors with six and nine elements, respectively, where some radiators in the outer periphery (three out of six sectors)

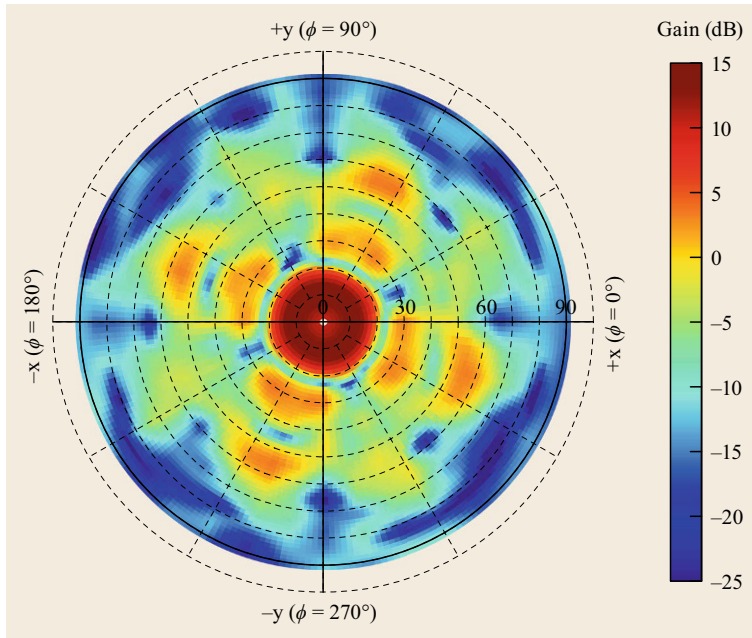


Fig. 17.32 Color-coded representation of the L1 gain pattern for a Block IIR-M antenna as a function of azimuth (ϕ) and boresight angle (after [17.64])

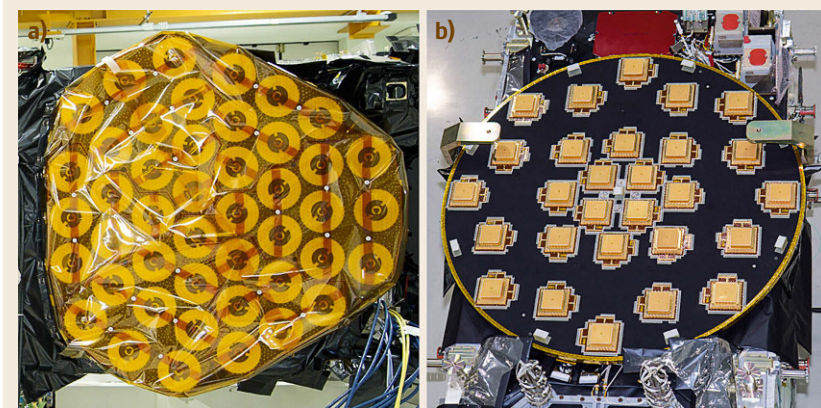


Fig. 17.33a,b Navigation antennas of the Galileo in-orbit validation (IOV) (a) and full operational capability (FOC) satellites (b) (courtesy of ESA and S. Corvaja (a) and OHB (b))

of the antenna are slightly shifted in order to get the required field distribution across the antenna aperture. The M-shaped beam pattern is obtained by feeding the central and the external part of the array in phase opposition. Moreover, sequential rotation has been applied between consecutive sectors to improve the cross-polar performance.

The radiator is an arrangement of stacked microstrip patches fed by a coaxial pin. The smaller patch on the top is a circular ring microstrip and radiates at the E1 band, while a larger circular patch at the bottom radiates at the E5 and E6 frequencies. The RHCP at both bands is achieved differently. The top ring has two notches, while the bottom patch uses orthogonal feeds with 90° to obtain circular polarization [17.65]. A circular metal-

lic wall surrounds the patches in order to shield the radiators in the array.

Key performance parameters of the IOV antenna are summarized in Table 17.3. Special care has also been taken to minimize passive intermodulation (PIM), that is, unwanted RF byproducts generated due to the nonlinear behavior of the transmitting system. Dedicated tests described in [17.66] confirm that PIM interference in the frequency bands of the near-by UHF search-and-rescue (SAR) antenna, the S-band telemetry, tracking and telecommand (TT&C) antenna as well as the C-band uplink receiver antenna are well below the noise and susceptibility level of those systems. Group delay variations in the individual frequency bands amount to $< 0.15\text{--}0.37$ ns (or $5\text{--}11$ cm), while PCVs are confined

Table 17.3 Galileo–IOV satellite antenna characteristics (after [17.65]) (LOC: limit-of-coverage, LB: low band, HB: high band)

Parameter	Value
Orbital position	medium altitude Earth orbit (MEO) (23 616 km)
Frequency plan	1145–1237 and 1259–1299 MHz (LB) 1555–1595 MHz (HB)
Polarization	RHCP both bands
Coverage and pattern	Isoflux corrected pattern (LOC: 12.67°) with 2 dB isoflux window
Minimum gain at LOC	15.35 dBi (HB), 14.85 dBi (LB)
Maximum AR	1.2 dB
Return loss and isolation	> 20 dB
Nominal power	103 W (LB), 75 W (HB)
Mass	< 15 kg
Envelope	1.38 m (diameter) 0.29 m (height)

to 4–7 mm over the 12.7° boresight angle range that covers the Earth from the orbital altitude of the Galileo satellites [17.67].

The Galileo FOC antenna (Fig. 17.33b) uses an array of stacked patch elements symmetrically arranged in four quadrants. It is inherited from the earlier GIOVE-A design but optimized to yield the desired performance with a lower number (28 versus 36) elements [17.68]. In addition, only the inner square of 12 elements covers the full set of three frequency bands (E1, E6, and E5), while the outer elements support either E1 and E6 (inner part of the outer ring) or E1 and E5 (outer part of outer ring). These design changes enabled a notably simplified feed network and helped to reduce the overall antenna mass to < 15 kg.

Compared to the fully populated IOV antenna, the arrangement of individual elements in the FOC antenna is similar to the helix antenna arrays discussed before and achieves the desired $J_1(x)/x$ intensity distribution in the aperture with a notably reduced number of patches. The M-shaped gain pattern of both Galileo antenna types is illustrated in Fig. 17.34 for the upper (1555–1595 MHz) and lower 1145–1299 MHz frequency bands. Both antennas closely approximate an ideal isoflux pattern with peak-to-peak gain variations of about 1.5–2.5 dB between nadir and the Earth’s limb.

17.5.3 Reflector-Backed Monofilar Antenna

As an alternative to the helix and patch antennas discussed before, the use of a reflector-backed monofi-

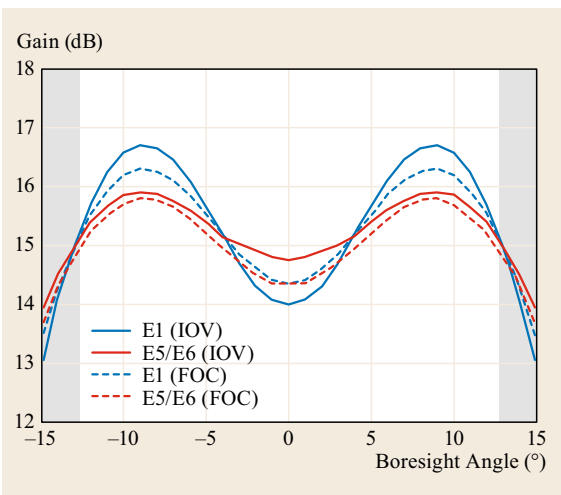


Fig. 17.34 Gain pattern of Galileo IOV and FOC satellites in the upper and lower frequency bands based (after [17.67, 68]). Shaded areas mark regions outside the apparent Earth disk

lar helix as GPS transmit antenna has been studied in [17.69]. In fact, a stepped parabolic reflector antenna had already been considered in the early design phase of the first GPS satellites. It offered a more uniform pattern than the helix array in use today but was discarded because of its notably higher side-lobes [17.61].

The design presented [17.69] uses a stack of two parabolic dishes with different diameter and focal length to achieve the desired M-shape gain pattern (Fig. 17.35). The monofilar helix is operated in back-fire mode, sending most of the radiation from the feed point toward the reflector. At a diameter of only 6 cm, the helix itself causes only marginal obstruction. The reported antenna gain ranges from 12–15 dBiC for different GNSS frequencies and initial tests confirm a high

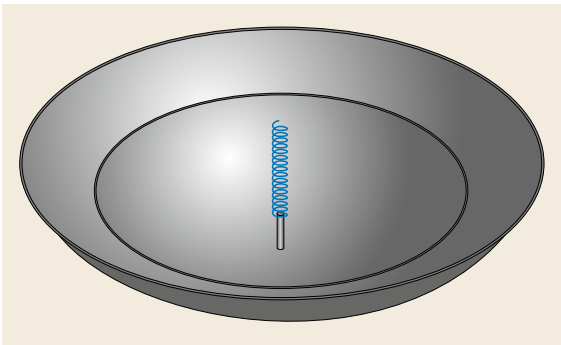


Fig. 17.35 GNSS transmit antenna design based on a back-fire monofilar helix element and a stacked reflector (after [17.69])

rotational symmetry of the achieved pattern. The simplicity of the overall design, which does not require an elaborate feed network, is claimed to result in no-

table cost savings compared to the traditional. However, reflector-based transmit antennas have not been adopted so far in actual GNSS satellites.

17.6 Antenna Measurement and Calibration

This final section discusses the testing and characterization of GNSS antennas. During the antenna design, the measurement of antenna parameters is of paramount importance to validate the desired function and performance. Although antenna simulation software gives the antenna engineer some idea of the expected behavior, human errors and manufacturing constraints may result in antennas which do not completely match the simulation results. Following a description of standard measurement techniques for the determination of basic antenna parameters, a dedicated subsection addresses the calibration of antenna phase patterns, which are of primary relevance for geodetic-grade user antennas as employed in all high-precision GNSS applications. Along with that, PC calibrations of GNSS transmit antennas are discussed.

17.6.1 Basic Antenna Testing

The testing of an antenna in general, and a GNSS antenna, in particular, requires sophisticated equipment for generating well-defined test signals and for measuring the overall system response. A typical setup for testing an antenna in receive mode comprises:

- A reference source antenna
- The transmit (TX) and receive (RX) system

- A positioning system
- The test facility (range).

These items are generic and are used for testing regardless of the antenna type or the subject application. Test setups for transmit antennas are similar to the above, except that the reference source antenna is replaced by a reference receive antenna. A sample facility for GNSS antenna testing is illustrated in Fig. 17.36.

The *reference source antenna* is the transmit antenna used to illuminate the antenna-under-test (AUT) and should have a stable PC and large bandwidth. Sectoral horns (linear polarized) and conical spirals (circularly polarized) are the few popular reference antennas. Since the GNSS antennas are designed to primarily receive or transmit RHCP signals, a conical spiral is usually the reference antenna of choice for the radiation pattern and gain measurements.

The *TX* and *RX* systems may comprise two distinct modules but can also be combined into a single unit. A vector network analyzer (VNA [17.71]) is a hybrid TX–RX test instrument very commonly used for the antenna gain and radiation pattern measurement. A VNA is used to measure four different *scattering parameters* (s-parameters) of a dual-port RF system such as a cable, filter, amplifier, or transmission line. With application

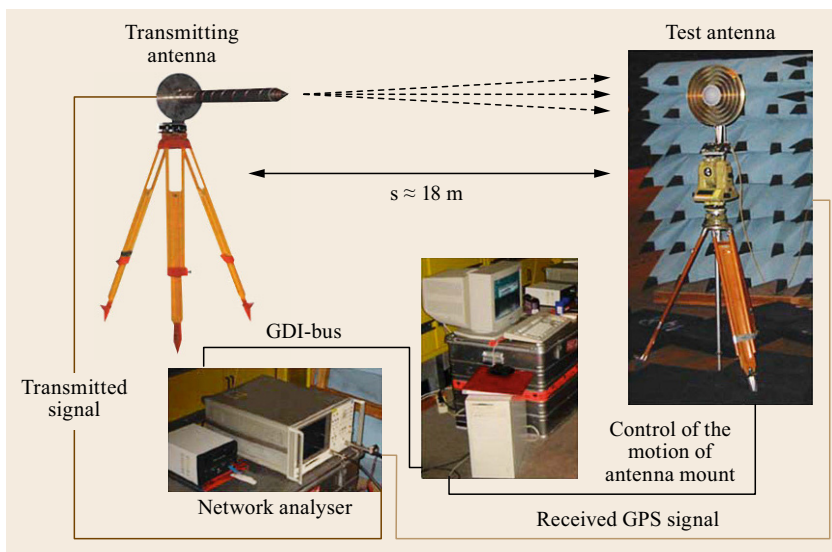


Fig. 17.36 Anechoic chamber measurement setup for antenna testing (after [17.70])

to antenna testing, these parameters have the following meaning:

- s_{11} , the reflection coefficient of the antenna connected at port 1
- s_{21} , the transmission (from port 1 to port 2) between the two antennas connected at the ports
- s_{12} , the transmission (from port 2 to port 1) between the two antennas connected at these ports
- s_{22} , the reflection coefficient of the antenna connected at port 2.

Based on these measurements, the desired parameters of the test antenna can later be derived.

The *positioning system* enables controlled orientation changes of the test antenna to assess the variation of antenna parameters with incidence angle. It comprises a gimbal mount or robotic arm with at least two rotation axes, which can be computer controlled to change the orientation (and, optionally position) of the antenna relative to the incident radiation

The last and final element in the antenna testing is the identification of the antenna *test facility*. The primary prerequisite for accurate antenna testing is that both the antennas should lie in the far field. This implies that the angular dependence of the transmit antenna field no longer varies with the distance and that the wavefronts received at the test antenna are sufficiently flat in the vicinity of the AUT.

Given the moderate size of common GNSS receive antennas and the employed signal wavelengths, the testing is usually conducted in a *anechoic chamber*, that is, a room that is capable of absorbing electromagnetic radiation. This anechoic chamber is made by lining up the walls, ceilings, and floor with special electromagnetic wave-absorbing material. The material is often cut in pyramids of various sizes so that any remaining reflections tend to spread in random directions and to add incoherently which further suppresses their impact on the antenna measurements.

Using the equipment described above, diverse antenna tests can be conducted to ensure that it fulfills design requirements. Similar to any other antenna, a GNSS antenna is tested for common parameters such as impedance matching and gain. However, a few specialized parameters are also tested to ensure that the antenna performance is suitable for GNSS applications. The parameters commonly measured in practical GNSS antenna tests include:

- Reflection coefficient and impedance matching
- Frequency of operation and bandwidth
- Radiation pattern
- Gain and efficiency

- Antenna polarization and AR, as well as other values introduced in Sect. 17.1.

The reflection coefficient, frequency of operation, and the bandwidth of the antenna can be obtained from the s_{11} reflection coefficient, when connecting the test antenna to the VNA and observing (s_{11} or s_{22}). The typical outcome of such a test is presented in Fig. 17.37, which shows the reflection coefficient of a dual-band step-shorted GNSS antenna. The dual-frequency capability can clearly be recognized from the two frequency dips. The same curve can also be used to assess the impedance bandwidth of the antenna.

The antenna gain and radiation pattern are obtained from the measurements of the s_{21} transmission coefficient across all incidence angles. However, appropriate calibration of the whole system is required in order to compensate for the transmission line as well as other losses. The gain pattern measurements can also be used to quantify the FBR as well as the multipath rejection ratio (Sect. 17.1) by comparing gains at different bore-sight angles.

Antenna polarization, finally, can be measured by testing the AUT against a reference antenna of known polarization. A receiving antenna will only have successful reception if the polarization matches that of the transmitting antenna. For quantitative AR measurements, the test antenna is measured successively against two linearly polarized antennas (vertical and horizontal).

17.6.2 Phase-Center Calibration

The performance achieved in carrier-phase-based positioning today (Chaps. 25 and 26) relies on high-accuracy observation models and a thorough character-

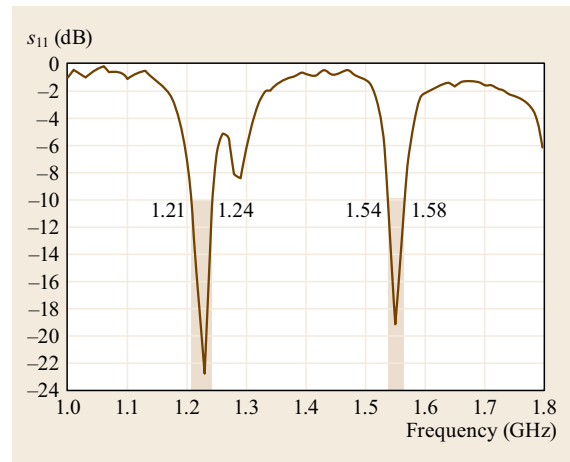


Fig. 17.37 s_{11} response of a dual-band GNSS antenna

ization of the measurement system. In order to exploit the (sub)millimeter precision of carrier-phase observations, a proper knowledge of the antenna PC is indispensable for both the receiver antenna and, in the case of absolute or long-baseline positioning, the transmit antennas onboard the GNSS satellites.

Driven by the needs of the geodetic community, dedicated antenna calibrations have, over many years, been conducted by various institutions in close cooperation with the international GNSS service (IGS; see Chap. 33). These calibrations presently cover a comprehensive set of geodetic-grade user antennas as well as the transmit antennas of GPS and GLONASS, which are sufficiently well observed by global GNSS monitoring stations [17.72].

Concepts and Conventions

As already discussed before, the phase center is considered as the point from which all radiated energy appears to emerge. Given the finite dimension and specific design of an antenna, its PC is not necessarily identical to the geometrical center and may even fall outside the actual structure. It is, therefore, common to specify the PCO of the antenna with respect to a mechanical *antenna reference point* (ARP) that is well accessible and can be easily related to a physical marker of a given survey point (Fig. 17.38). For user antennas, the PCO is typically specified in a coordinate system aligned with the boresight–symmetry axis of the antenna (z-axis) and a fixed axis in the antenna plane (y-axis) identified by a north marker. Based on the assumption of a perfect antenna setup the (x, y, z)-axes of the antenna system are commonly equated to the

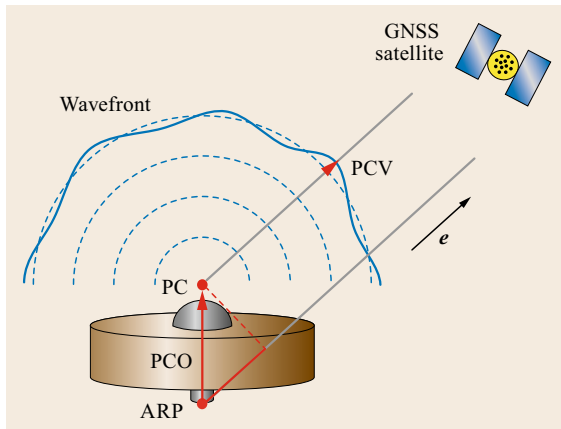


Fig. 17.38 Schematic illustration of the antenna reference point (ARP), the phase center (PC) and phase center offset (PCO), as well as the PCVs for a receiver antenna. The dotted lines indicate idealized spherical wavefronts around the average PC

east, north, and up direction when specifying antenna PCOs.

For GNSS satellite antennas, in contrast, a coordinate system aligned with the main body axes of the satellite is adapted, where the z-axis denotes again the (Earth-pointing) boresight direction, while the y-axis is aligned with the rotation axis of the solar panels. Details of the axes conventions adopted by the International GNSS service (IGS) for each constellation and type of satellites are documented in [17.73]. For practical reasons, GNSS antenna PCOs are commonly referred to the center of mass (CoM) of the spacecraft.

However, the concept of a unique PC is based on the idealized assumption of spherical wavefronts with a common center. In reality, the wavefront suffers various distortions induced by asymmetries of the antenna elements and the influence of the supporting structure on which the antenna is mounted (Fig. 17.38). The associated deviations from a point-like PC model are described by additional phase range corrections known as PCVs. Compared to the nominal range between the receiver ARP and the GNSS satellite CoM, the observed range differs by

$$\zeta_{\text{PCO/PCV}} = \mathbf{e}^\top (\mathbf{r}_{\text{PCO,sat}} - \mathbf{r}_{\text{PCO,rcv}}) + (\zeta_{\text{PCV,sat}}(-\mathbf{e}) + \zeta_{\text{PCV,rcv}}(\mathbf{e})), \quad (17.16)$$

where \mathbf{e} denotes the receiver-to-satellite unit vector, \mathbf{r}_{PCO} is the PCO vector (expressed in the same coordinate system), and ζ_{PCV} describes the direction-dependent PCO for the satellite antenna (index *sat*) and the receiving antenna (index *rcv*), respectively. Equation (17.16) provides the basic observation model (Chap. 19) for the determination of PCO/PCV values from actual GNSS measurements as well as their application in precise positioning applications.

It may be recognized that (17.16) does not provide a unique distinction between PCOs and PCVs, since a change in the PCO value can always be compensated by a corresponding PCV change, a normalizing condition is required to separate both effects in an unambiguous manner. Common conventions include a zero-mean condition for the PCV over a given range of boresight angles or a zero PCV along the boresight axis.

Antenna calibration values recommended for geodetic processing and compatible with IGS orbit and clock products are distributed as part of the IGS antenna model [17.72], which is continuously updated to include new antennas and satellites. The respective PCO and PCV values are distributed in a standardized ANTenna EXchange format (ANTEX, see [17.74] and Annex A), which facilitates a consistent application in the positioning and orbit determination software.

Receiver Antenna Calibration

Common techniques for the calibration of receiving antennas comprise of anechoic chamber measurements as well as relative and absolute field calibrations [17.75]. All of these can provide PC and pattern information for individual frequencies, but differ in a variety of practical aspects.

Anechoic chamber measurements, as described in the previous section, are among the earliest techniques employed for the PC characterization of geodetic antennas [17.70, 76]. Here, an artificial signal source – ideally a GNSS signal simulator – is used to provide navigation signals, and the phase response is measured while the test antenna is rotated around two independent axes. Anechoic chamber measurements, so far, provide the only means for PCO/PCV determination across all frequency bands utilized by current and emerging navigation satellite systems [17.77], but are so far limited in their availability.

Given this limitation and the complexity of anechoic chamber phase pattern measurements, relative field calibration has early-on been established as an alternative technique for antenna calibrations [17.78] and applied, among others, by the National Geodetic Survey (NGS) to characterize the phase patterns of most geodetic antennas. In this method, the receiver antenna is calibrated relative to a standard reference antenna. Both antennas are set up in close proximity on well-surveyed points and their relative position is determined from differential carrier-phase observations to obtain the PCO. Phase residuals with respect to the computed PC are then used to derive the PCVs.

To overcome the limitations of relative calibrations in the analysis of global geodetic networks, the IGS moved to an absolute antenna model in November 2006 [17.79]. It is based on field calibrations for the user antennas, which are obtained from GNSS measurements collected with a robot-mounted antenna (Fig. 17.39). Other than for a static antenna, the robot enables rapid changes of the boresight direction and the rotation angle. In this way, carrier-phase observations with a smooth hemispherical coverage of the antenna field of view can be collected in a short test campaign, based on which the PCO and PCVs can be inferred [17.80, 81]. Similar to chamber calibrations, the robotic measurements rely on the highly accurate knowledge of the translational motion of the ARP during antenna orientation changes.

In view of their availability for a comprehensive set of antennas as well as due to small differences between chamber and robot calibrations [17.70] that are a subject of ongoing research, only robot calibrations as well as converted relative field calibrations are currently incorporated into the IGS antenna model, which so far



Fig. 17.39 Robotic test stand of Geo++, Hannover, for absolute field calibration of GNSS antennas (after [17.70])

(igs08.atx [17.72]) covers the L1 and L2 frequencies for GPS and GLONASS. An example PCV from robot calibrations of a geodetic antenna in the L1 frequency band is shown in Fig. 17.14. Benefits of the absolute PC model for the analysis of geodetic time series and the terrestrial reference frame are discussed in [17.79, 82].

Satellite Antenna Calibration

Similar to receiver antennas, the aforementioned field and chamber calibration methods can likewise be applied to determine the PCO and/or PCVs of transmit antennas onboard the GNSS satellites. By way of example, relative field and robot calibration have been performed for the GPS Block II/IIA antenna array [17.83, 84], while PC/pattern calibrations using anechoic chambers or outdoor far-range test stands have, for example, been conducted for the antennas of the GPS Block IIR/IIR-M satellites [17.64, 85], selected antennas of the Galileo program [17.67, 86], as well as the two L-band antennas of the QZS-1 satellite. While these measurements can deliver PCO and/or PCV information for individual signals frequencies, they have neither been consistently applied to all GNSS satellites

nor can they fully account for the impact of the satellite body on the radiation field.

As part of the IGS antenna model, a different calibration technique is therefore employed, which estimates the antenna PCOs and, optionally, variations of all relevant GNSS satellites from observations collected with a global network of monitoring stations [17.87, 88]. Using the previously established absolute phase pattern corrections for the receiver antennas, the satellite PCOs (and PCVs) as defined in (17.15) can be obtained by adjusting these values along with other parameters, such as satellite orbits, station coordinates, and Earth orientation parameters from the observations (Chap. 34). The determination of PCOs for GPS and GLONASS satellites using this approach is discussed in [17.79, 89, 90] and a compilation of current values as used in the igs08.atx antenna model is given in [17.72]. For the BeiDou MEO and inclined geo-synchronous orbit (IGSO) satellites, initial PCO estimates have been reported in [17.91, 92], but further refinement will be required to establish a consolidated set of values for use in the IGS antenna model. For the Galileo FOC and IOV satellites, a first set of PCOs derived from observations of the IGS multi-GNSS network is reported in [17.93].

In various cases [17.89, 94, 95], azimuth-dependent PCVs have been obtained along with the PCOs of the GNSS transmit antennas. These patterns exhibit representative peak-to-peak amplitudes at the 10–30 mm

level and typically reflect the symmetry properties of the respective antenna arrays. For practical purposes, only the boresight angle variation of the GNSS transmit antenna pattern is presently taken into account in the IGS antenna model. Even though the consideration of azimuth-dependent PCVs is likely to enable a further improvement in the GNSS data processing, they are not presently taken into account due to a lack of consistent PCV calibrations from multiple IGS analysis centers.

Other than chamber calibrations, the transmit antenna PCOs and PCVs cannot be determined for individual signal frequencies within the GNSS network processing but refer to a specific ionosphere-free linear combination of observations collected on two different frequencies (e.g., L1/L2 for GPS and GLONASS, E1/E5a for Galileo and B1/B2 for BeiDou). This constitutes an obvious limitation of the approach for future multifrequency applications and will need to be considered in the evolution of the IGS antenna models.

Acknowledgments. The authors of this chapter would like to thank the Editors of the book for their guidance during the development of this chapter. The authors would also like to thank Prof. Steffen Schön, Lori Winkler, and Prof. Dmitry Tatarnikov, who took time from their busy schedules and helped in getting the clearance for some of the key images required for this chapter.

References

- 17.1 J.D. Krauss, R.J. Marhefka: *Antennas for all Applications*, 3rd edn. (McGraw Hill Higher Education, Columbus 2003)
- 17.2 C.A. Balanis: *Antenna Theory: Analysis and Design*, 3rd edn. (Wiley, Chichester 2005)
- 17.3 B.R. Rao: Introduction to GNSS antenna performance parameters. In: *GPS/GNSS Antennas*, ed. by B.R. Rao, W. Kunysz, R. Fante, K. McDonald (Artech House, Boston 2013), Chap. 1–62
- 17.4 G.J.K. Moernaut, D. Orban: GNSS antennas – An introduction to bandwidth, gain pattern, polarization, and all that, *GPS World* **20**(2), 42–48 (2009)
- 17.5 B.R. Rao: FRPAs and high-gain directional antennas. In: *GPS/GNSS Antennas*, ed. by B.R. Rao, W. Kunysz, R. Fante, K. McDonald (Artech House, Boston 2013) pp. 119–133, Chap. 2
- 17.6 X. Chen, C.G. Parini, B. Collins, Y. Yao, M.U. Rehman: *Antennas for Global Navigation Satellite Systems* (Wiley, Chichester 2012)
- 17.7 D. Orban, G.J.K. Moernaut: *The Basics of Patch Antennas* (Orban Microwave, Orlando 2009)
- 17.8 E.O. Hammerstad: Equations for microstrip circuit design, *Proc. 5th Eur. Microw. Conf.*, Hamburg (IEEE) (1975) pp. 268–272
- 17.9 C.C. Chen, S. Gao, M. Maqsood: Antennas for global navigation satellite system receivers. In: *Space Antenna Handbook*, ed. by W.A. Imbriale, S. Gao, L. Boccia (Wiley, Chichester 2012) pp. 548–595, Chap. 14
- 17.10 H. Chen, K. Wong: On the circular polarization operation of annular-ring microstrip antennas, *IEEE Trans. Antennas Propag.* **47**(8), 1289–1292 (1999)
- 17.11 J.M. Tranquilla, S.R. Best: A study of the quadrifilar helix antenna for global positioning system (GPS) applications, *IEEE Trans. Antennas Propag.* **38**(10), 1545–1550 (1990)
- 17.12 P.K. Shumaker, C.H. Ho, K.B. Smith: Printed half-wavelength quadrifilar helix antenna for GPS marine applications, *Electron. Lett.* **32**(3), 153–153 (1996)
- 17.13 C.C. Kilgus: Resonant quadrifilar helix design, *The Microw. J.* **13**(12), 49–54 (1970)
- 17.14 D. Lamensdorf, M. Smolinski: Dual band quadrifilar helix antenna, *Proc. IEEE Antennas Propag. Soc. Int. Symp.*, San Antonio (2002) pp. 488–491
- 17.15 P.G. Elliot, E.N. Rosario, R.J. Davis: Novel quadrifilar helix antenna combining GNSS, iridium, and a UHF communications monopole, *Proc. Mil. Commun. Conf.*, Orlando (2012) pp. 1–6

- 17.16 S. Liu, Q.-X. Chu: A novel dielectrically-loaded antenna for tri-band GPS applications, Proc. 38th European Microw. Conf., Amsterdam (2008) pp. 1759–1762
- 17.17 F. Scire-Scappuzzo, S.N. Makarov: A low-multipath wideband GPS antenna with cutoff or non-cutoff corrugated ground plane, IEEE Trans. Antennas Propag. **57**(1), 33–46 (2009)
- 17.18 NovAtel Inc.: GPS-704X Antenna Design and Performance (Calgary 2013) <http://www.novatel.com/assets/Documents/Papers/GPS-704xWhitePaper.pdf>
- 17.19 W. Kunysz: High Performance GPS Pinwheel Antenna, Proc. ION GPS 2000, Salt Lake City, UT 19–22 Sep. 2000 (ION, Virginia 2000) 2506–2511
- 17.20 J.J.H. Wang: Antennas for global navigation satellite system (GNSS), Proc. IEEE **100**(7), 2349–2355 (2012)
- 17.21 GPS world antenna survey 2014, GPS World **25**(2), 37–57 (2014)
- 17.22 E. Levine: A review of GPS antennas, Consum. Electron. Times **3**(3), 233–241 (2014)
- 17.23 R. Schmid: igs08.atx antenna model, <https://igs.cb.jpl.nasa.gov/igs08/station/general/igs08.atx>
- 17.24 Environmental Conditions and Test Procedures for Airborne Equipment, RTCA DO-160G Change 1 (RTCA, Washington DC 2014)
- 17.25 Minimum Operational Performance Standards for Airborne Supplemental Navigation Equipment Using Global Positioning System (GPS), RTCA DO-208, 07/12/1991 (RTCA, Washington DC 1991)
- 17.26 O. Montenbruck, M. Garcia-Fernandez, Y. Yoon, S. Schön, A. Jäggi: Antenna phase center calibration for precise positioning of LEO satellites, GPS Solutions **13**(1), 23–34 (2009)
- 17.27 J. Wettergren, M. Bonnedal, P. Ingvarson, B. Wästberg: Antenna for precise orbit determination, Acta Astronaut. **65**(11), 1765–1771 (2009)
- 17.28 M. Ohgren, M. Bonnedal, P. Ingvarson: GNSS antenna for precise orbit determination including SIC interference predictions, Proc. 5th EUCAP, Rome (2011) pp. 1990–1994
- 17.29 J. Santiago-Prowald: L. Drioli Salghetti: Space environment and materials. In: *Space Antenna Handbook*, Chap. Vol. 4, ed. by W.A. Imbriale, S. Gao, L. Boccia (Wiley, Chichester 2012) pp. 106–132
- 17.30 R.L. Fante, K.F. McDonald: Adaptive GPS antennas. In: *GPS/GNSS Antennas*, ed. by B.R. Rao, W. Kunysz, R. Fante, K. McDonald (Artech House, Boston 2013), Chap. 1–62
- 17.31 D. Reynolds, A. Brown, A. Reynolds: Miniaturized GPS Antenna Array Technology and Predicted Anti-Jam Performance, Proc. ION GPS 1999, Nashville, TN 14–17 Sep. 1999 (ION, Virginia 1999) 777–786
- 17.32 M.M. Casabona, M.W. Rosen: Discussion of GPS anti-jam technology, GPS Solutions **2**(3), 18–23 (1999)
- 17.33 M.V.T. Heckler, M. Cuntz, A. Konovaltsev, L. Greda, A. Dreher, M. Meurer: Development of robust safety-of-life navigation receivers, microwave theory and techniques, IEEE Trans. **59**(4), 998–1005 (2011)
- 17.34 A. Hornbostel, N. Basta, M. Sgammini, L. Kurz, S.I. Butt, A. Dreher: Experimental results of interferer suppression with a compact antenna array, Proc. ENC-GNSS 2014, Rotterdam (ENC, 2014) pp. 1–14
- 17.35 W. Kunysz: Advanced Pinwheel Compact Controlled Reception Pattern Antenna (AP-CRPA) designed for Interference and Multipath Mitigation, Proc. ION GPS 2001, Salt Lake City, UT 11–14 Sep. 2001 (ION, Virginia 2001) 2030–2036
- 17.36 W.C. Cheuk, D.A. Trinkle, M. Gray: Null-steering LMS dual-polarised adaptive antenna arrays for GPS, J. Global Position. Syst. **4**(1/2), 258–267 (2005)
- 17.37 K. Wu, L. Zhang, Z. Shen, B. Zheng: An anti-jamming 5-element GPS antenna array using phase-only nulling, Proc. 6th Int. Conf. ITS Telecommun., Chengdu 2006, ed. by G. Wen, S. Komaki, P. Fan, G. Landrac (2006) pp. 370–373
- 17.38 S. Jin, E. Cardellach, F. Xie: *GNSS Remote Sensing* (Springer, Dordrecht 2014)
- 17.39 M. Unwin, S. Gleason, M. Brennan: The Space GPS Reflectometry Experiment on the UK Disaster Monitoring Constellation Satellite, Proc. ION GPS 2003, Portland, OR 9–12 Sep. 2003 (ION, Virginia 2003) 2656–2663
- 17.40 M. Unwin, S. Gao, R. De Vos Van Steenwijk, P. Jales, M. Maqsood, C. Gommenginger, J. Rose, C. Mitchell, K. Partington: Development of low-cost spaceborne multi-frequency GNSS receiver for navigation and GNSS remote sensing, Int. J. Space Sci. Eng. **1**(1), 20–50 (2013)
- 17.41 G. Foti, C. Gommenginger, P. Jales, M. Unwin, A. Shaw, C. Robertson, J. Roselló: Spaceborne GNSS reflectometry for ocean winds: First results from the UK TechDemoSat-1 mission, Geophys. Res. Lett. **42**(13), 5435–5441 (2015)
- 17.42 C. Rocken, R. Anthes, M. Exner, D. Hunt, S. Sokolovskiy, R. Ware, M. Gorbunov, W. Schreiner, D. Feng, B. Herman, Y.-H. Kuo, X. Zou: Analysis and validation of GPS/MET data in the neutral atmosphere, J. Geophys. Res. **102**, 29849–29866 (1997)
- 17.43 P. Silvestrin, R. Bagge, M. Bonnedal, A. Carlstrom, J. Christensen, M. Hagg, T. Lindgren, F. Zangerl: Spaceborne GNSS Radio Occultation Instrumentation for Operational Applications, ION GPS 2000, Salt Lake City, UT 19–22 Sep. 2000 (ION, Virginia 2000) 872–880
- 17.44 P. Ingvarson: GPS receive antennas on satellites for precision orbit determination and meteorology, Proc. 17th Int. Conf. Microw., Radar Wirel. Commun., MIKON 2008, Wroclaw (2008) pp. 1–6
- 17.45 C.C. Counselman: Multipath-rejecting GPS antennas, Proc. IEEE **87**(1), 86–91 (1999)
- 17.46 V. Filippov, D. Tatarnicov, J. Ashjaee, A. Astakhov, I. Sutiagin: The First Dual-Depth Dual-Frequency Choke Ring, Proc. ION GPS 1998, Nashville, TN 15–18 Sep. 1998 (ION, Virginia 1998) 1035–1040
- 17.47 M. Maqsood, S. Gao, T. Brown, M. Unwin, R.D. Steenwijk, J.D. Xu: A compact multipath mitigating ground plane for multiband GNSS antennas, IEEE Trans. Antennas Propag. **61**(5), 2775–2782 (2013)

- 17.48 L. Boccia, G. Amendola, G. Di Massa, L. Giulichi: Shorted annular patch antennas for multipath rejection in GPS-based attitude determination systems, *Microw. Opt. Technol. Lett.* **28**(1), 47–51 (2001)
- 17.49 L.I. Basilio, R.L. Chen, J.T. Williams, D.R. Jackson: A new planar dual-band GPS antenna designed for reduced susceptibility to low-angle multipath, *IEEE Trans. Antennas Propag.* **55**(8), 2358–2366 (2007)
- 17.50 S.F. Mahmoud, A.R. Al-Ajmi: A novel microstrip patch antenna with reduced surface wave excitation, *Prog. Electromag. Res.* **86**, 71–86 (2008)
- 17.51 L. Boccia, G. Amendola, S. Gao, C. Chen: Quantitative evaluation of multipath rejection capabilities of GNSS antennas, *GPS Solutions* **18**(2), 199–208 (2014)
- 17.52 D. Tatarnikov, A. Astakhov, A. Stepanenko: Convex GNSS reference station antenna, *Proc. Int. Conf. Multimedia Technol., Hangzhou 2011 (ICMT 2011)* pp. 6288–6291
- 17.53 D. Tatarnikov, A. Astakhov, A. Stepanenko: Broad-band convex impedance ground planes for multi-system GNSS reference station antennas, *GPS Solutions* **15**(2), 101–108 (2011)
- 17.54 W. Kunysz: A three dimensional choke ring ground plane antenna, *Proc. ION GPS 2003, Portland, OR 9–12 Sep. 2003 (ION, Virginia 2003)* 1883–1888
- 17.55 L. Bedford, N. Brown, J. Walford: New 3D Four Constellation High Performance Wideband Choke Ring Antenna, *ION ITM 2009, Anaheim, CA 26–28 Jan. 2009 (ION, Virginia 2009)* 829–835
- 17.56 J. Walford: *State of The Art, Leading Edge Geodetic Antennas from Leica Geosystems – Leica Reference Antennas White Paper* (Leica Geosystems, Heerbrugg 2009)
- 17.57 M. Maqsood, S. Gao, T. Brown, J.D. Xu, J.Z. Li: Novel multipath mitigating ground planes for multiband global navigation satellite system antennas, *Proc. 6th EUCAP 2012, Prague (2012)* pp. 1920–1924
- 17.58 Y. Rahmat-Samii, F. Yang: *Electromagnetic Band Gap Structures in Antenna Engineering* (Cambridge Univ. Press, Cambridge 2009)
- 17.59 R. Baggen, M. Martinez-Vazquez, J. Leiss, S. Holzwarth, L.S. Drioli, P. de Maagt: Low Profile Galileo antenna using EBG technology, *IEEE Trans. Antennas Propag.* **56**(3), 667–674 (2008)
- 17.60 J.S. Ajioka, H.E. Harry Jr: Shaped beam antenna for earth coverage from a stabilized satellite, *IEEE Trans. Antennas Propag.* **18**(3), 323–327 (1970)
- 17.61 F.M. Czopek, S. Shollenberger: Description and performance of the GPS Block I and II L-Band antenna and link budget, *Proc. ION GPS 1993, Salt Lake City, UT 22–24 Sep. 1993 (ION, Virginia 1993)* 37–43
- 17.62 C. Brumbaugh, A.W. Love, G. Randall, D. Waineo, S.H. Wong: Shaped beam antenna for the global positioning satellite system, *Proc. Antennas Propag. Soc. Int. Symp., Amherst (1976)* pp. 117–120
- 17.63 F.H. Bauer: GNSS space service volume and space user data update, *Proc. 10th Meet. Int. Commun. GNSS (ICG), Working Group A, Boulder (UNOOSA, Vienna 2015)* pp. 1–34
- 17.64 W. Marquis: The GPS Block IIR/IIR-M antenna panel pattern. Lockheed Martin Corp. (2014) <http://www.lockheedmartin.com/us/products/gps/gps-publications.html>
- 17.65 A. Montesano, C. Montesano, R. Caballero, M. Naranjo, F. Monjas, L.E. Cuesta, P. Zorrilla, L. Martinez: Galileo system navigation antenna for global positioning, *Proc. 2nd EuCAP 2007, Edinburgh (IET, Stevenage 2007)* pp. 1–6
- 17.66 F. Monjas, A. Montesano, C. Montesano, J.J. Llorente, L.E. Cuesta, M. Naranjo, S. Arenas, I. Madrazo, L. Martínez: Test campaign of the IOV (in orbit validation) Galileo system navigation antenna for global positioning, *Proc. 4th EUCAP 2010, Barcelona (2010)* pp. 1–5
- 17.67 F. Monjas, A. Montesano, S. Arenas: Group delay performances of Galileo system navigation antenna for global positioning, *Proc. 32nd ESA Antenna Workshop on Antennas for Space Applications, Noordwijk (2010)* pp. 1–8
- 17.68 P. Valle, A. Netti, M. Zolesi, R. Mizzone, M. Bandinelli, R. Guidi: Efficient dual-band planar array suitable to Galileo, *Proc. 1st EUCAP 2006, Nice (2006)* pp. 1–7
- 17.69 Y.U. Kim: An M-shaped beam producing dual stacked reflector antenna for GPS satellite applications, *Proc. IEEE Int. Symp. Antennas Propag., San Diego (2008)* pp. 1–4
- 17.70 B. Görres, J. Campbell, M. Becker, M. Siemes: Absolute calibration of GPS antennas: Laboratory results and comparison with field and robot techniques, *GPS Solutions* **10**(2), 136–145 (2006)
- 17.71 Understanding the Fundamental Principles of Vector Network Analysis, Application Note 1287-1 (Agilent, Santa Clara 2012)
- 17.72 R. Schmid, R. Dach, X. Collilieux, A. Jäggi, M. Schmitz, F. Dilssner: Absolute IGS antenna phase center model igs08.atx: Status and potential improvements, *J. Geodesy* **90**(4), 343–364 (2016)
- 17.73 O. Montenbruck, R. Schmid, F. Mercier, P. Steigenberger, C. Noll, R. Fatkulin, S. Kogure, S. Ganeshan: GNSS satellite geometry and attitude models, *Adv. Space Res.* **56**(6), 1015–1029 (2015)
- 17.74 M. Rothacher, R. Schmid: ANTEX: The Antenna Exchange Format, Version 1.4, 15 Sep. 2010 <ftp://igs.org/pub/station/general/antex14.txt>
- 17.75 M. Rothacher: Comparison of absolute and relative antenna phase center variations, *GPS Solutions* **4**(4), 55–60 (2001)
- 17.76 B.R. Schupler: The response of GPS antennas – How design, environment and frequency affect what you see, *Phys. Chem. Earth, Part A* **26**(6–8), 605–611 (2001)
- 17.77 M. Becker, P. Zeimet, E. Schönmann: Anechoic chamber calibrations of phase center variations for new and existing GNSS signals and potential impacts in IGS processing, *Proc. IGS Workshop 2010, Newcastle upon Tyne (IGS, Pasadena 2010)*, pp. 1–44, 28 June–2 July 2010
- 17.78 G.L. Mader: GPS antenna calibration at the National Geodetic Survey, *GPS Solutions* **3**(1), 50–58 (1999)
- 17.79 R. Schmid, P. Steigenberger, G. Gendt, M. Ge, M. Rothacher: Generation of a consistent absolute phase center correction model for GPS receiver and

- satellite antennas, *J. Geodesy* **81**(12), 781–798 (2007)
- 17.80 G. Wübbena, M. Schmitz, F. Menge, V. Böder, G. Seeber: Automated Absolute Field Calibration of GPS Antennas in Real-Time, *Proc. ION GPS 2000*, Salt Lake City, UT 19–22 Sep. 2000 (ION, Virginia 2000) 2512–2522
- 17.81 M. Schmitz, G. Wübbena, G. Boettcher: Tests of phase center variations of various GPS antennas, and some results, *GPS Solutions* **6**(1/2), 18–27 (2002)
- 17.82 P. Steigenberger, M. Rothacher, R. Schmid, A. Rülke, M. Fritsche, R. Dietrich, V. Tesmer: Effects of different antenna phase center models on GPS-derived reference frames. In: *Geodetic Reference Frames, IAG Symposia 134*, ed. by H. Drewes (Springer, Heidelberg 2009) pp. 83–88
- 17.83 G.L. Mader, F. Czopek: Calibrating the L1 and L2 phase centers of a Block IIA antenna, *Proc. ION GPS 2001*, Salt Lake City, UT 11–14 Sep. 2001 (ION, Virginia 2001) 1979–1984
- 17.84 G. Wübbena, M. Schmitz, G. Mader, F. Czopek: GPS Block II/IIA Satellite Antenna Testing using the Automated Absolute Field Calibration with Robot, *Proc. ION GNSS 2007*, Fort Worth, TX 25–28 Sep. 2007 (ION, Virginia 2007) 1236–1243
- 17.85 W.A. Marquis, D.L. Reigh: The GPS block IIR and IIR-M broadcast L-band antenna panel: Its pattern and performance, *Navigation* **62**(4), 329–347 (2015)
- 17.86 R. Zandbergen, D. Navarro: Specification of Galileo and GIOVE Space Segment Properties Relevant for Satellite Laser Ranging ESA-EUING-TN/10206 (ESA/ESTEC, Noordwijk 2008) iss. 3.2
- 17.87 R. Schmid, M. Rothacher: Estimation of elevation-dependent satellite antenna phase center variations of GPS satellites, *J. Geodesy* **77**(7/8), 440–446 (2003)
- 17.88 R. Schmid, M. Rothacher, D. Thaller, P. Steigenberger: Absolute phase center corrections of satellite and receiver antennas, *GPS Solutions* **9**(4), 283–293 (2005)
- 17.89 F. Dilssner, T. Springer, C. Flohrer, J. Dow: Estimation of phase center corrections for GLONASS-M satellite antennas, *J. Geodesy* **84**(8), 467–480 (2010)
- 17.90 R. Dach, R. Schmid, M. Schmitz, D. Thaller, S. Schaer, S. Lutz, P. Steigenberger, G. Wübbena, G. Beutler: Improved antenna phase center models for GLONASS, *GPS Solutions* **15**(1), 49–65 (2011)
- 17.91 F. Dilssner, T. Springer, E. Schönmann, W. Enderle: Estimation of satellite antenna phase center corrections for BeiDou, *Proc. IGS Workshop 2014*, Pasadena (IGS, Pasadena 2014) p. 1
- 17.92 J. Guo, X. Xu, Q. Zhao, J. Liu: Precise orbit determination for quad-constellation satellites at Wuhan University: Strategy, result validation, and comparison, *J. Geodesy* **90**(2), 143–159 (2016)
- 17.93 P. Steigenberger, M. Fritsche, R. Dach, R. Schmid, O. Montenbruck, M. Uhlemann, L. Prange: Estimation of satellite antenna phase center offsets for Galileo, *J. Geodesy* **90**(8), 773–785 (2016)
- 17.94 B.J. Haines, Y.E. Bar-Sever, W. Bertiger, S. Desai, N. Harvey, J. Weiss: Improved models of the GPS satellite antenna phase and group-delay variations using data from low-earth orbiters, *Proc. AGU Fall Meet. 2010*, San Francisco (AGU, Washington DC 2010) pp. 1–12
- 17.95 F. Dilssner: GPS IIF-1 satellite – Antenna phase center and attitude modeling, *Inside GNSS* **5**(6), 59–64 (2010)

18. Simulators and Test Equipment

Mark G. Petovello, James T. Curran

This chapter presents a review of a range of global navigation satellite system (GNSS) simulators and test equipment. Different types of systems are discussed, including radio frequency (RF) and intermediate frequency (IF) simulators; record and playback systems; and measurement simulators. The key features of each of these devices are examined, illustrating their various implementations, typical usage, and highlighting their individual benefits and drawbacks. The chapter concludes with an overview of considerations that should be borne in mind when selecting and using simulators and test equipment.

18.1	Background	537	18.3	IF-Level Simulators	546
18.1.1	Received RF Signal	537	18.3.1	Implementation	547
18.1.2	GNSS Receivers	540	18.3.2	Important Considerations	548
18.1.3	GNSS Simulators	540	18.4	Record and Playback Systems	549
18.1.4	Record and Playback Systems	542	18.4.1	Implementation	550
18.1.5	Details	543	18.4.2	Important Considerations	550
18.2	RF-Level Simulators	543	18.5	Measurement-Level Simulators	552
18.2.1	Implementation	544	18.5.1	Implementation	553
18.2.2	Important Considerations	544	18.5.2	Important Considerations	553
			18.6	Combining Live and Simulated Data	554
			18.6.1	Implementation	555
			18.6.2	Important Considerations	555
			18.7	Other Considerations	556
			18.7.1	GNSS Systems Supported	556
			18.7.2	Interference and Spoofing	556
			18.7.3	Other Data	556
			18.7.4	Configurability	556
			18.7.5	Expandability	556
			18.8	Summary	557
			References		557

As a complement to previous chapters on receiver design and position estimation, this chapter looks at the role of global navigation satellite system (GNSS) signal simulators and related test equipment in the development of GNSS receivers.

Simulation plays many roles in various stages of a receiver's lifespan: initial component selection, algorithm development, fault-isolation and debugging, performance assessment, and final production-line quality-control, to name but a few. While all of these tasks could be performed using live GNSS signals, the use of a simulation resource can significantly simplify the task, increase efficiency, and provide higher test fidelity.

To generalize somewhat, one can consider that there are three types of testing that are conducted on re-

ceivers: known-response tests, performance tests, and exploratory tests.

Known-response testing is typical of quality control and fault-isolation, wherein the receiver is subject to a specific and well-defined scenario or stimulus, and its response is compared to a well-known reference response, representative of a properly functioning receiver. For quality control, a threshold is typically defined in terms of the difference between the reference and measured responses, and a pass or fail verdict can be ascribed. For fault-isolation or debugging tests, the particular stimulus may be carefully chosen to excite a specific feature or functionality of the receiver, thereby enabling a user to identify where the fault may lie.

Performance testing often takes a similar form, whereby a receiver may be subject to a well-defined

scenario or stimulus, and one or many of a set of typical receiver parameters may be examined. Common metrics describing receiver performance include, for example, time-to-first fix, acquisition or tracking thresholds, raw measurement accuracy, or position, velocity, and time (PVT) accuracy. These tests may be deterministic or statistical and may either be absolute measurements or may be relative to some reference data provided as input to, or generated by, the simulator. This kind of testing is often conducted, for example, during algorithm development, receiver prototyping or data-sheet generation.

Exploratory testing is the set of experiments or tests that an engineer may conduct to examine the behavior of a receiver or a system in a new or non-nominal scenario in order to evaluate new or different receiver algorithms or architectures. These tests are generally more qualitative than quantitative and seek to understand the behavior of the receiver, or to compare the behavior of multiple receivers under different scenarios.

Although these types of tests are quite different in nature and purpose, they have a number of common requirements: they require signals or stimuli that have the appearance of having originated from genuine GNSS transmissions and they require well-defined reference data. Genuine GNSS signals, of course, can be employed to excite or stimulate the receiver in the desired fashion, but may require a difficult, lengthy, and/or laborious test campaign. For example, inducing high-signal dynamics may require the use of a specialized vehicle, while inducing transmission channel effects may require dedicated test ranges. Moreover, attaining accurate reference data for the test can prove quite difficult. In these cases, the use of a simulator can greatly simplify and enhance the test procedure.

Unlike live testing, simulators generate highly controllable, configurable, and repeatable inputs to GNSS receivers. Moreover, depending on the part of the receiver that is to be tested, different types of simulators may be used. The most common type of simulator generates a radio frequency (RF) signal, but simulators can also be used to generate intermediate frequency (IF) samples, for injection behind the RF down-conversion stage. Simulators that generate receiver measurements can also be developed, which are useful for assessing navigation algorithms. All types of simulators allow

users to choose from a range of constellation configurations, from a single, stationary satellite, to one or more systems (the U.S. Global Positioning System GPS, the Russian GLObal NAVigation Satellite System, GLONASS, the European Galileo, the Chinese BeiDou, etc.), at a single frequency or at multiple frequencies.

Depending on its purpose or format, a simulator may reproduce a version of a genuine scenario with varying degrees of fidelity. The simulated scenario need only reflect the genuine scenario with sufficient accuracy so as not to influence the results of a given test. Basic quality-control testing of a specific receiver component or function may only require a loosely modeled GNSS signal, perhaps as simple as a single-channel simulator producing a carrier modulated by the ranging codes and data. In contrast, to test the full navigation capabilities of a receiver, it may be necessary to employ a simulator producing a full constellation of satellite signals, including relevant errors and propagation phenomena. Testing receivers for mobile cellular handsets may require accurate modeling of a land-mobile satellite channel, whereas testing surveying receivers may place more emphasis on atmospheric errors. If only the baseband component of the receiver is to be tested, then a simulator may produce signals as they would appear at the entry to a baseband processor in a genuine GNSS receiver.

An interesting tradeoff between live testing and simulation is the use of record and playback systems. Unlike simulators, record and playback systems collect genuine GNSS data and allow the receiver developer to replay it as many times as necessary. These systems provide similar repeatability to that of traditional simulation tools, but make an interesting tradeoff between configurability and fidelity: capturing genuine GNSS signals ensures that all of the features of the genuine signals including all propagation phenomena are accurately captured. However, the exact properties of the signal that is captured may not be fully known and so acquiring accurate reference data may be difficult.

This chapter addresses all of the above aspects. It begins with an overview of GNSS and GNSS receivers and then discusses the basic requirements of a simulator. Different types of simulators and simulation techniques are introduced, discussed, and contrasted.

18.1 Background

This section provides an overview of GNSS and GNSS receivers from a simulation perspective. Equations are developed to describe the received signal as well as intermediate steps within a GNSS receiver. Based on this, different types of simulators are briefly defined, as are record and playback systems. The different types of simulators are then addressed in detail in later sections.

For the purpose of this chapter, Fig. 18.1 is a convenient high-level breakdown of GNSS and GNSS receivers. The left side of the figure represents the physical world and consists of two main parts. First, the physics and geometry component involves the position and velocity of the receiver and all satellites in view. This is used to compute the geometric range and range rate between the receiver and each satellite which is then contaminated by satellite and atmospheric errors. The second component represents the propagation channel which varies significantly between applications. In general, however, it accounts for effects such as free-space loss (geometric spreading), multipath, fading, shadowing, and antenna effects.

On the right side of Fig. 18.1 is a breakdown of a GNSS receiver under test. The first component is the front-end which is responsible for mixing the RF signal to near baseband, filtering, and sampling. This is also

the first place where the receiver's oscillator is used. Second, the signal-processing component correlates the locally generated and incoming signals for use in signal acquisition and tracking. The outputs of this component are the pseudorange (code phase), carrier-phase, and carrier Doppler measurements that are passed to the third component of the receiver, namely the navigation solution. The navigation solution is responsible for generating the PVT estimates of the receiver.

18.1.1 Received RF Signal

This section presents a mathematical description of the RF signal at the output of a receiver's antenna (i. e., the left side of Fig. 18.1).

Physics and Geometry

The primary concern of the physics and geometry component is to describe the relationship between the receiver and a satellite, including all of the systematic errors related to propagation. In reality, satellites blindly broadcast signals toward the Earth, synchronized as closely as possible to their corresponding system time, and physics takes care of the rest. At the receiver, the satellite signals that are observed at any

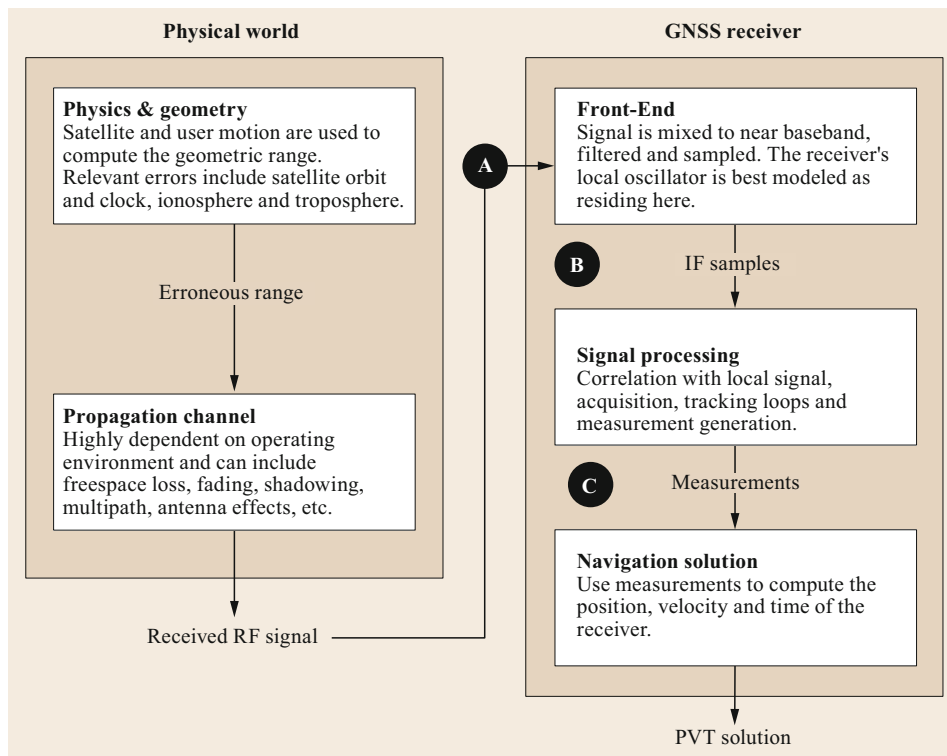


Fig. 18.1 Breakdown of GNSS from a simulation perspective. The left side is a high-level description of what is taking place in the physical world that results in a signal observed at the receiver's antenna. The right side shows a high-level breakdown of a GNSS receiver; point A is the input for RF simulators, point B is the input for IF simulators, and point C is the input for measurement simulators

given point in time, say t_R , originated from some previous point in time, say t_T , from wherever the satellite was at t_T . Herein, t_R is the time of reception or *receive time*, and t_T is the time of transmission or *transmit time*. The time of flight, Δt_{TOF} , is thus the difference between the receive and transmit times

$$\Delta t_{\text{TOF}} = t_R - t_T. \quad (18.1)$$

For navigation, a receiver must compute the range between where the satellite *was* at t_T and where the receiver *is* at t_R , all within the Earth-centered, Earth-fixed coordinate frame at the receive time. Since the receiver operates in its own time frame, t_R is readily available even if it is nominally biased, and t_T can be extracted from the satellite's navigation message. The pseudorange is then resolved as the difference between t_R and t_T scaled by the speed of light.

GNSS simulators are fundamentally different than what is described above because they are receiver-centric. That is, at a high level, input to a simulation consists of a series of positions and times of a *receiver* as well as a set of equations for the satellite positions over time (ephemerides), and it is the responsibility of the simulator to determine the transmit times for each satellite and, by extension, their corresponding position and velocity at that time.

Considering, for simplicity, the line-of-sight (LOS) case, the signal arriving at the receiver is effectively the same as the signal that left the satellite Δt_{TOF} time ago. Thus, if $s_T(t)$ is the signal that left the satellite, then the signal that arrives at the receiver, $s_R(t)$, is approximately given by (a more detailed equation is given in (18.8))

$$s_R(t_R) \approx s_T(t_R - \Delta t_{\text{TOF}}). \quad (18.2)$$

So the first task of a simulator is to calculate Δt_{TOF} by examining how long it takes a signal to propagate from the satellite to the receiver. This is approximated here by the time needed to travel the geometric range at the speed of light in vacuum, plus some atmospheric-related delays

$$\Delta t_{\text{TOF}} = \frac{1}{c} \|\mathbf{r}(t_R) - \mathbf{r}_{\text{Sat}}(t_T)\| + \frac{1}{c} L_{\text{atm}}, \quad (18.3)$$

where $\mathbf{r}(t_R)$ is the receiver position vector at time t_R , $\mathbf{r}_{\text{Sat}}(t_T)$ is the satellite position vector at time t_T , and L_{atm} represents the delay induced by the atmosphere, in units of length, including the ionosphere and troposphere. Solving (18.1) for t_T and substituting the result into (18.3) yields an expression for Δt_{TOF} as an explicit

function of t_R

$$\Delta t_{\text{TOF}}(t_R) = \frac{1}{c} \|\mathbf{r}(t_R) - \mathbf{r}_{\text{Sat}}(t_R - \Delta t_{\text{TOF}}(t_R))\| + \frac{1}{c} L_{\text{atm}}. \quad (18.4)$$

Of interest is that the first term on the right side of this equation is only a function of the inputs to the simulator, namely receive time, receiver position, and satellite position (ephemeris). The second term needs to be modeled by the simulator.

Unfortunately, even if one ignores the atmospheric effects, (18.4) is somewhat difficult to handle as Δt_{TOF} appears on both sides. Moreover, as has been discussed in Chap. 3, the function describing the satellite position is not simple since it must account for, amongst other things, the effect of Earth rotation during Δt_{TOF} [18.1]. Fortunately, (18.4) is quite smooth and can be readily solved using recursion.

It is also noted that by changing t_R , (18.4) implicitly accounts for Doppler effects. Specifically, realizing that the numerator of the first term on the right-hand side of (18.4) is the geometric range, the derivative of Δt_{TOF} with respect to t_R is, to first order, given by

$$\frac{d\Delta t_{\text{TOF}}}{dt_R} = \frac{1}{c} \dot{\rho} + \frac{1}{c} \frac{dL_{\text{atm}}}{dt_R}, \quad (18.5)$$

where $\dot{\rho}$ is the geometric range rate which is related to the perceived Doppler shift by dividing by the carrier wavelength. The range rate can be expressed in terms of the satellite and receiver trajectories at any given instant by [18.2]

$$\dot{\rho} = \frac{(\mathbf{r}(t_R) - \mathbf{r}_{\text{Sat}}(t_T)) \cdot (\mathbf{v}(t_R) - \mathbf{v}_{\text{Sat}}(t_T))}{\|\mathbf{r}(t_R) - \mathbf{r}_{\text{Sat}}(t_T)\|}, \quad (18.6)$$

where $\mathbf{v}_{\text{Sat}}(t_T)$ is the satellite velocity at the time of transmission and $\mathbf{v}(t_R)$ is the receiver velocity at the time of reception. This expression is quite useful as $\dot{\rho}$ actually changes relatively slowly with time, and is approximately constant over a millisecond, that is, the smallest ranging code period used in GNSS (Chaps. 7–10). Simulators often exploit this fact and perform a linear approximation to the Δt_{TOF} function with piece-wise sections of constant $\dot{\rho}$, effectively producing signals which step between short periods of constant Doppler.

Before writing the equation for the received signal, a simplified equation for the signal transmitted by the i -th satellite can be written as

$$s_{T,i}(t) = AC_i(t)D_i(t) \cos(2\pi f_{\text{Nom}}t), \quad (18.7)$$

where A is the amplitude of the transmitted signal which is assumed constant over time and across satellites, C is the ranging code sequence, D is the navigation data bit sequence, and f_{Nom} is the nominal RF carrier frequency (e.g., 1575.42 MHz for GPS L1 C/A). Equation (18.7) assumes that only a single signal on a single carrier is broadcast, but this can be easily generalized by adding additional terms. Similarly, different types of ranging codes and pilot signals can also be accommodated with additional terms. Then, the composite received signal can be expressed as the sum of the signals that left the satellites at their respective times of transmission

$$s_R(t_R) = \sum_i^{N_{SV}} \alpha_i(t_R) s_{T,i}(t_R - \Delta t_{\text{TOF},i} + dt_{\text{SV},i}) + n(t_R), \quad (18.8)$$

where N_{SV} is the number of satellites in view of the receiver, α is the attenuation of the received signal due to the propagation channel, and $n(t)$ is thermal noise. These latter two terms are expressed as a function of receive time because they are primarily a function of receiver location and/or hardware. An astute reader will have also noticed the inclusion of one extra term, $dt_{\text{SV},i}$ representing the satellite clock error. At the time the signal left the satellite, it was modulated according to the satellite's onboard clock. Although the true time may have been t_T , according to the onboard clock, it is $dt_{\text{SV},i}$ seconds later/earlier, and so the signal parameters – code phase, carrier-phase, carrier Doppler, and data bit – will actually represent that at time $t_T + dt_{\text{SV},i}$. Thus, satellite clock errors manifest themselves as range errors, just as receiver clock biases do, albeit at the *other* end of the signal.

Finally, it is noted that while the above model represents the LOS case only, a similar model can be readily constructed for the specular multipath case, provided a suitable equivalent for (18.3) can be found that represents the geometry of a reflected signal.

Propagation Channel

The propagation channel nominally encompasses all of the effects on the signal between the satellite and the receiver. Having already discussed the factors influencing the time of flight, this section examines the factors affecting the signal quality. The key effects to be considered are briefly discussed below.

Free-Space Loss. Also called geometric spreading, this accounts for the loss of power intensity per unit area as the signal propagates away from the satellite. For Earth-based GNSS applications, the free-space loss differs by at most about 1 dB (depending on satellite elevation), but this variability is largely reduced [18.3] by

proper shaping of the satellite's antenna gain pattern. In contrast, space-based applications may experience very different received powers due to the much wider range of distances to the satellites and the satellite antenna gain pattern (discussed below) [18.4–6].

Shadowing. This is the attenuation of signals resulting from propagation through natural or man-made obstructions such as trees, people, buildings, or vehicles [18.7, 8]. Depending on the receiver, shadowing may prohibit tracking of the signal.

Antenna Effects. With reference to Chap. 17, the most important antenna effects in the context of simulation are its gain pattern and axial ratio, which affect the received signal power. For Earth-based applications, the primary concern is for the receiver antenna, including phase variations that result from changing the orientation of the antenna relative to the satellites due to phase wind-up [18.9, 10] or phase-center variation, which, in turn, may depend on the assumed dynamics of the receiver (e.g., when a ship turns, the antenna also rolls). However, for space-based applications, the effect of the satellite's antenna gain pattern is critically important for correctly determining the received signal power [18.4–6].

High-accuracy positioning applications are also concerned with the phase-center variation/stability of the antenna which can introduce small (millimeter- to centimeter-level) systematic or randomly varying errors [18.11]. Some users may also be concerned with the antenna frequency response, perhaps wishing to investigate the rejection of out-of-band interference.

Multipath and Fading. As discussed in Chap. 15, multipath arises when a signal arrives at the receiver via multiple paths. Multipath is primarily a function of the reflectors in the vicinity of the receiver, but the large number of parameters needed to model the reflectors (i.e., their electrical characteristics, locations relative to the receiver, etc.) make it virtually impossible to predict in real time. A negative by-product of multipath is the decrease in received signal power that results from the destructive interference of two or more signals at the antenna. In extreme cases and/or when the signal is already attenuated, these *fades* may cause the receiver to temporarily lose lock on the signal [18.7, 8].

After the consideration of the propagation channel, the signal received via a single path still satisfies the general form of (18.8) with two differences. First, the amplitude term is the combination of several effects

$$\alpha(t_R) = \alpha_{\text{FSL}}(t_R) \cdot \alpha_{\text{Shdw}}(t_R) \cdot \alpha_{\text{Ant}}(t_R), \quad (18.9)$$

where α_{FSL} is the attenuation due to free-space loss, α_{Shdw} is the attenuation due to shadowing, and α_{Ant} encapsulates antenna gain pattern effects (amplification or attenuation). Second, in the presence of multipath, all received paths need to be included in the summation in (18.8).

18.1.2 GNSS Receivers

This section provides a high-level description of what happens within different components of the receiver. In particular, features of the received signal that have a significant impact on different receiver components are highlighted. This will be used in Sect. 18.1.3 to motivate different types of simulators and their features.

Front-End

The front-end is generally the interface between the received RF signal (either real or simulated) and the remainder of the receiver. Its purpose is to accept an RF signal from the receive antenna or directly from the simulator, and produce digitized IF samples for the signal-processing stage. A receiver's front-end is generally comprised of a signal pre-amplification and filtering stage, one or more stages of down-conversion, anti-aliasing filters, gain control and, finally, analog-to-digital conversion (ADC; see Chap. 13).

The key factors that are of interest in measuring the performance of the front-end include: the oscillator quality as observed in the IF signal phase process; noise performance or noise figure; spectrum quality, including appropriate suppression/rejection of local oscillator harmonics; gain control, and digitizer performance. When available, direct analysis of the output of the front-end provides a useful means of assessing these factors.

Signal Processing

The receiver's signal processing is discussed in detail in Chap. 14. Here, it is noted that the basic task conducted in the signal-processing stage of a receiver is to process IF samples in order to produce pseudorange (code phase), carrier-phase, and carrier Doppler measurements for the navigation solution. Of particular interest are the parts of the received signal that stress the acquisition and tracking algorithms. For example, the received signal and noise levels are particularly significant in the acquisition stages; multipath and fading are generally most significant in the code-tracking algorithms; and, reference oscillator stability, scintillation, and LOS dynamics have the most impact on carrier-phase tracking [18.12–14].

In most cases, with the exception of vector or ultra-tightly (deeply) coupled receivers (Chaps. 13 and 18), the signal-processing stage operates on a signal-by-signal basis. From a tracking perspective, the loop bandwidths are generally wide enough such that slowly-varying errors due to atmospheric delays, orbital errors, and intersystem clock biases, to name but a few, are not observable here. Practically, this means that these errors pass through this part of the receiver unchanged and thus only manifest themselves in the measurements. More details on the outputs of the signal-processing stage are as described in Chap. 19.

Navigation Solution

The navigation solution accepts the pseudorange, carrier-phase and/or carrier Doppler measurements generated by the signal-processing step and uses them to compute the PVT solution of the receiver. The key challenge here is to remove as many of the systematic errors as possible prior to processing the measurements. Blunder detection (data snooping; see Chap. 24) should also be included in order to improve the overall reliability of the solution. Details for computing the PVT solution are included in later chapters.

18.1.3 GNSS Simulators

Having broken down GNSS and GNSS receivers in Fig. 18.1, this section introduces different types of simulators, their key requirements, and their purpose. Details of different simulators are given in Sects. 18.2–18.5.

Types of Simulators

Recalling Fig. 18.1, points labeled A, B, and C represent different locations within a receiver where simulated data can be injected. By extension, depending on what part of the receiver is being tested, points B and C, plus the PVT solution itself, can be interpreted as outputs that can be analyzed to assess performance. With this in mind, three types of simulators are identified:

- **RF-Level Simulators:** Using a combination of software and hardware, these simulators generate RF signals that can be fed directly into any commercial GNSS receiver. RF simulators allow for complete end-to-end testing of the receiver with the exception of the antenna, whose effects are usually modeled in the simulator itself such that the signal does not need to be broadcast over the air; this is called *conductive testing*. However, rebroadcasting the RF signal in an anechoic chamber would allow the antenna to be included in the testing chain;

this is called *rebroadcast testing*. Correspondingly, RF-level simulators are the most common types of simulator. Current manufacturers of such simulators, and sample current models, include: Spirent's GSS9000; Spectracom's GSG-6 Series; Rohde and Schwarz's SMBV100A Vector Signal Generator; and IFEN's NAVX-NCS product line.

- **IF-Level Simulators:** By generating a sequence of IF samples, IF simulators bypass the receiver's front-end and provide input directly to the signal-processing stage. Since most commercial receivers do not provide direct access to the signal-processing stage, IF simulators are typically reserved for research and in-house testing and/or used for testing software receivers – where all digital processing is performed using a reconfigurable architecture, be it pure-software, reconfigurable hardware such as field programmable gate array (FPGAs), digital signal processors (DSPs), or some combination thereof. IF simulators can, however, be coupled with a playback system (Sect. 18.1.4), to effect an RF simulator. A number of IF simulator products are currently available (most including playback capability as well), including: Avera's URT-5000; National Instruments' Global Navigation Satellite System Toolkits; M3 Systems' GNSS Test Platform; and RACELOGIC's LabSat 3 GPS Simulator.
- **Measurement-Level Simulators:** These simulators are used for testing a receiver's navigation solution, or an independent data-processing software. Assuming a receiver can properly track a signal without introducing any unexpected effects, measurement simulators are primarily concerned with modeling the systematic GNSS errors (e.g., orbit errors, atmospheric effects, etc.) as well as the stochastic errors arising from noise and multipath. Several different measurement-level simulators have been produced, many of them based on MATLAB®. Some of the more common simulators are the SATNAV Toolbox 3.0 from GPSoft LLC, the GPS Toolbox 5 from L3NAV Systems and the Galileo Service Volume Simulator from the European Space Agency (ESA) [18.15].

The type of simulator affects two things: the stages of the receiver that can be tested and the phenomena that the simulator needs to simulate. Obviously, a simulator can only be used to test components of the receiver downstream of its entry point. By extension, for a given entry point, the simulator will usually simulate all of the *relevant* errors upstream from the entry point to adequately satisfy the testing requirements. For example, IF simulators can only test a receiver from its

signal-processing stage and later, in which case the simulator must simulate the effects of the front-end and earlier.

The concept of *relevant* errors is critically important and can vary based on what is being evaluated. If the user intends to perform an *end-to-end* test starting from a given entry point through to the PVT solution, then indeed all effects above the entry point should be accounted for in the simulation. However, some tests may focus on a single stage of the receiver that may ease the simulation requirements. For example, consider testing only the signal-processing stage. As discussed in Sect. 18.1.2, low-frequency errors such as orbit error, atmospheric errors, and so on have no appreciable effect on this stage and thus would not need to be simulated. In contrast, tests of the navigation solution will almost certainly require simulating the low-frequency errors.

In general, simulating every effect is ideal, but this may be too costly and/or too time consuming, especially if a simulator is being developed in-house. In such cases, the user must decide if certain errors can be omitted from the simulation, or perhaps simulated with lower fidelity. This decision will depend on the type of simulator available and on the level of access to intermediate outputs within the receiver.

Key Requirements

Regardless of the type of simulator considered and the tests being conducted, all simulators operate on the idea of simulating a particular *scenario*. Scenarios may differ for a variety of reasons including the location of the user, type of propagation channels considered, assumed receiver dynamics, level of atmospheric errors, etc. In all cases, however, simulators should provide scenarios that are:

- **High-Fidelity:** The simulator should be capable of producing a signal that is consistent with what the receiver is expecting and what, specifically, is being tested. This includes signal fidelity (structure, bandwidth, etc.), error fidelity, motion fidelity, etc. Some of these aspects, such as user motion, can be replicated with very high accuracy. Other errors, shadowing, for example, cannot be modeled easily or reliably; the model may be too smooth, not smooth enough, or may not exhibit the full range of values. Specifications for the fidelity of the simulation may vary considerably between application, receiver under test, and/or what part of the receiver is being tested.
- **Controllable:** The user should be able to modify/configure scenario parameters to excite the re-

ceiver in an appropriate manner. A simulator should be configurable to produce a sufficiently wide variety of scenarios so as to adequately evaluate receiver performance. The more variety that can be provided, the better the features and limitations of a receiver can be identified.

- **Repeatable:** A simulator should be capable of reproducing a scenario with sufficiently high repeatability that successive runs of an experiment produce equivalent results to within the expected level of accuracy. Repeatability is important when a user wishes to make precise measurements and requires that the test/retest variability is low. Otherwise, poor repeatability will result in high experiment variability and may mask the measurement being made. This can be particularly important when analyzing chaotic phenomena, such as carrier-phase cycle-slipping, receiver loss-of-lock and re-acquisition behavior, where small differences in the initial conditions can have a very large effect on the experiment outcome. The level of repeatability of a particular test will also have a direct impact on the precision with which quality control can be conducted.

18.1.4 Record and Playback Systems

Closely related to simulators is the concept of record and playback. Figure 18.2 shows a high-level view of a testing scenario involving a record and playback system. The *recording* phase accepts live RF signals as input, uses a front-end to shift the signal to IF, samples and digitizes the signal, and then writes the IF samples to a file.

Then, at some later point in time, the *playback* phase reads the IF samples and converts them to an analog signal at IF, shifts the signal back up to RF, and then outputs the RF signal.

The key difference of a record and playback system relative to a simulator is the loss of control over the input. This can be viewed as both a positive and negative characteristic. On the one hand, without special data-collection processes or equipment, it prevents the determination of the receiver's absolute position and the propagation/measurement errors that were present, thus limiting the ability to fully test the receiver. On the other hand, it allows for testing under *real* operating environments and, as such, circumvents any limitations that may exist in the simulator models.

In light of the above, record and playback systems can be viewed as hybrids of RF-level and IF-level simulators. Similar to an RF simulator, they generate an RF signal that can be used to test all aspects of the receiver. They even provide an additional advantage of includ-

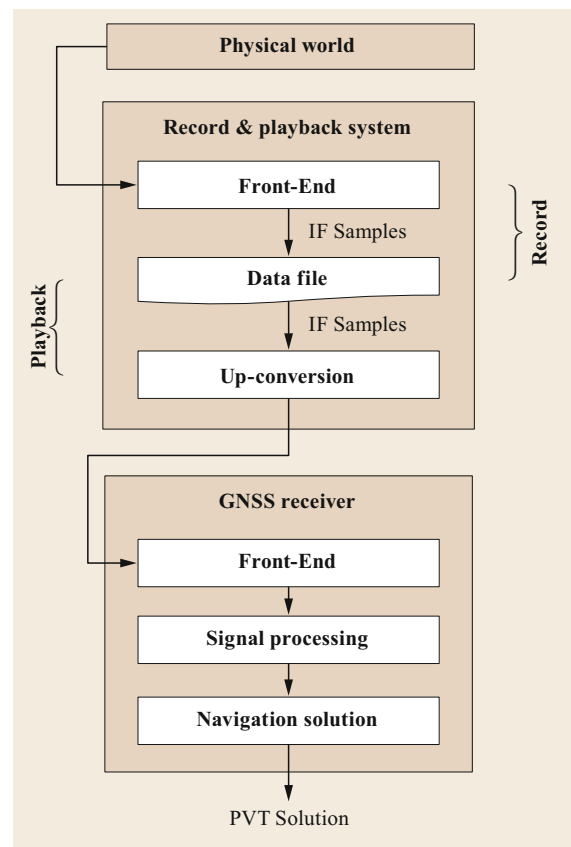


Fig. 18.2 Flowchart for a record and playback system. Although drawn as a continuous process, the recording and replay are done as separate steps

ing real antenna effects that were incurred during the recording phase. At the same time, the IF samples could be fed directly into a receiver, bypassing the RF portion altogether, thus mimicking an IF simulator. Conversely, if an IF simulator is available external to the record and playback system, the record and playback system could be used to convert those IF samples to RF (as discussed in Section 18.2, this is becoming the norm).

Record and playback systems can also be used in innovative ways to improve or expand testing capabilities. For example, strong signals could be collected, but the replayed signal could be attenuated before passing the RF signal to the receiver. Similarly, interference-free data could be collected and different types of interferences could be added after the fact to assess its effect on the receiver [18.16, 17]. Or, since the recording phase is not limited to GNSS signals (except perhaps in terms of the center frequency and bandwidth), the system could be used to collect interference signals that occur in select locations and then replayed and combined with real or simulated GNSS signals.

Although relatively new to the GNSS community, several record and playback systems are commercially available. Examples include Avera's RP-3200 Wideband RF Record & Playback system, RACELOGIC's LabSat 3 GPS Simulator, and Spirent's GSS6400. Other systems, such as the National Instruments USRP-based (Universal Software Radio Peripheral) system, can also be used for GNSS with the added benefit of being able to also record and playback signals in other frequency bands.

18.2 RF-Level Simulators

An RF GNSS simulator presents to the receiver an RF feed similar to what it would expect at the output of an antenna. In the case that the receiver has an enclosed or nonremovable antenna, it is possible to rebroadcast the simulated RF signal using a transmit antenna and a suitable anechoic chamber or enclosure to avoid multipath effects. Being the earliest entry point to the receiver flow, RF simulation maximizes the number of features of the receiver that can be examined.

The choice of simulator is generally constrained by the tests that are to be carried out, the features of the receiver the user wishes to examine, and by the accessibility of the receiver. For example, an RF simulator is appropriate when the user wishes to:

- Conduct an (almost; the antenna is usually omitted) end-to-end test of a receiver under predefined and repeatable scenarios, for example, when developing receiver specifications (acquisition and tracking) or in a quality-control procedure.
- Test the front-end in isolation, when IF samples are accessible to the tester, such that the performance

18.1.5 Details

The rest of this chapter expands on each of the different GNSS simulators introduced above, including record and playback systems. Emphasis is placed upon implementation details, important considerations and, where appropriate, practical challenges and issues helpful for those with less experience with GNSS receiver testing. After each type of simulator is presented, the combination of real and simulated signals is discussed which can prove to be a powerful testing approach.

of the filtering, down-conversion, and digitization stages can be examined.

- Examine the performance of the signal-processing and navigation solution algorithms when the only accessible input point to the receiver is the RF stage.
- Examine the coexistence or interoperability of other systems such as terrestrial communication networks with GNSS, or examine the effects of various interferences. In such cases, it may not be possible or practical to simulate these other systems/signals at the IF or measurement level.

An example of a bench-top RF simulator and accompanying shielding chamber is shown in Fig. 18.3. The simulator depicted may be used alone to perform conductive testing of entire receivers or receiver components. Alternatively, it can be coupled with the shielding chamber to perform either conductive tests, or broadcast testing of devices with integrated antennas. The chamber is designed to shield the device under test from external interferences, and absorb signals radiated within the chamber to minimize reflections.



Fig. 18.3 An example of a bench-top simulator and accompanying RF shielding chamber suitable for conductive or rebroadcast testing of GNSS receivers or receiver components

In other cases, where it is desirable to perform a broadcast test, such that the effects of the antenna pattern are represented, it is necessary to place the system under test in the far-field of the broadcast antenna, which for GNSS requires a separation of some meters. For example, it may be of interest to observe receiver operation when the antenna observes GNSS signals from a high elevation and some interference signals from low or negative elevations. Such a configuration is illustrated in Fig. 18.4.

18.2.1 Implementation

RF GNSS simulators have changed quite significantly in the last decade. What were originally very large, expensive, and power-hungry devices are now available in a wide range of sizes and match a wide range of budgets. Such simulators exist as large cabinet-style installations, as bench-top units and even as hand-held devices, of course, with varying degrees of capability and accuracy. Although this technology is changing quite rapidly, RF simulators still, typically, fall into two main categories: those which dedicate a single hardware channel to each simulated signal, in some cases including one additional channel per multipath reflection; and those which synthesize all signals in the digital domain and perform a collective digital-to-analog conversion (DAC) and up-conversion to RF.

Those that dedicate a single channel to each simulated signal typically offer systems within the range of 16–64 channels per unit. Generally, the number of channels is extended further by combining/connecting more than one unit into a single simulator. Such sim-

ulators utilize essentially the same technology as the satellite transmitters themselves and have dominated the market until recently, as the technology to synthesize signals digitally in real time was unavailable or had limited capability. Examples of this kind of simulator include, for example, some units from IFEN and Spirent.

Recently, however, this has changed and modern simulation techniques begin with the digital synthesis of multiple signals, followed by collective DAC and up-conversion to RF. Examples of such, so-called, *software-defined* simulators include those from Rohde and Schwarz, Spirent and Spectracom. In fact, today, there are many real-time software-defined simulators available such as, for example, Avera's URT-5000, National Instruments' Global Navigation Satellite System Toolkits, M3 Systems' GNSS Test Platform, and RACELOGIC's LabSat systems. Software-defined digital synthesis has the distinct advantage that there is no intrinsic limitation to the number of signals that can be simulated (including multipath signals), an important consideration for those working with multi-frequency and/or multi-GNSS receivers. They also offer easier upgrading capabilities if additional features and/or systems need to be added later.

18.2.2 Important Considerations

It is important to consider that the first gain stage essentially defines the noise figure of the system. Therefore, when performing a rebroadcasting test on a receiver that has an integrated antenna or when an active antenna must be included in the test chain, it is preferable

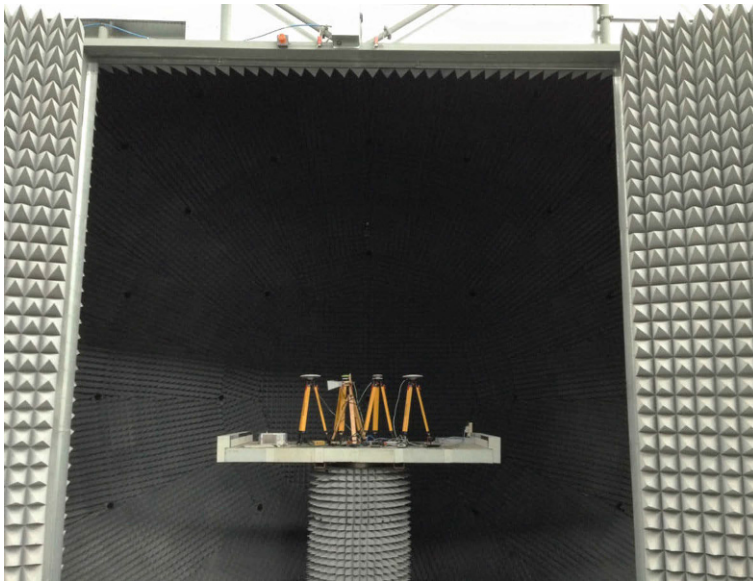


Fig. 18.4 An anechoic chamber located at the Joint Research Center, Italy, during a GNSS interference test. The receivers and antennas under test are mounted on a central platform while simulated GNSS signals are broadcast from a fixed overhead antenna and a selection of interference signals are broadcast from a mobile antenna, from a variety of elevations

to simulate a well-calibrated signal without noise. In this way, the noise performance of the active receiver elements are reflected in the test results, with the carrier power being controlled by the simulator and the noise contribution being produced in the receiver itself. Accurately calibrating this broadcast power may not always be possible, however. When the broadcast signal cannot be accurately calibrated in absolute terms, it is better to broadcast both the signal and a simulated noise, such that a relative carrier-to-noise floor can be accurately synthesized, albeit at a noncalibrated absolute power level. This rebroadcast signal must be sufficiently higher than the true thermal noise floor of the receiver, so as to preserve the simulated carrier-to-noise ratio. Care must be taken, however, to keep the broadcast power within the levels expected by the receiver.

Similarly, when performing a conductive test, the receiver RF chain will have a specific noise temperature and will introduce into the system a certain level of nonrepeatable additive noise. As tests are repeated, this thermal noise will differ and will have a corresponding impact on the overall receiver performance. To reduce the effect of this issue to a certain extent, simulators can synthesize both simulated signals at a simulated thermal noise, having the appropriate relative power levels, but both at a higher power level than the true thermal noise floor of the receiver.

This issue of additive noise also plays into the repeatability of certain tests. RF simulators cannot provide an exactly repeatable test scenario. Repeatability of RF simulators is influenced, primarily, by three factors: thermal noise, reference oscillators, and start-time synchronization. By simulating both signal and noise, as mentioned previously, the contribution of the true thermal noise can be reduced, leaving only the repeatable simulated component.

The stability of the simulator's reference oscillator can also play a critical role. Any asynchronicity between the simulator and receiver timescales will appear as apparent receiver clock errors, regardless of which clock is more stable or better calibrated. To minimize this effect, it is important that the simulator clock be more stable than that of the receiver. The nature of GNSS navigation poses relatively strict requirements on the stability and turn-on bias of the oscillators used in receivers (at least relative to other telecommunication systems), and so RF simulators must be equipped with high-quality oscillators.

Nonetheless, the instability of the simulator's reference oscillator can influence the repeatability of tests. Crystal oscillators are affected by aging, temperature, physical orientation, and physical stress, all of which will influence the nominal frequency of the oscillator,

and they also exhibit stochastic instability which produces a random variation in the oscillator output over time [18.18–20]. These effects are present on both the simulator and the receiver side and will lead to variability in the overall experiment. Variability arises due to the simulator and the receiver time frames moving relative to one another, and respectively relative to the true passage of time.

In the case that it is not necessary to include the receiver clock in the test path, the issue of oscillator instability can be partially circumvented by providing both the simulator and the receiver with a common clock or a common pulse per second (PPS), or by providing either the simulator or receiver with a clock or PPS from the other. Indeed, in many cases, an RF simulator will be capable of providing or accepting both a reference clock and a PPS signal, although it is not always the case that the receiver under test can accept or provide either of these. If it is the case that the quality of the receiver clock is the object of the test, then sharing a common clock or PPS is not an option and, therefore, it must be ensured that the simulator reference clock is at least an order of magnitude more stable than that of the receiver. This fact can impose real limitations on what level of testing can be conducted with a simulator. In practice, this problem is generally addressed by providing the simulators with a GNSS-disciplined reference clock exhibiting good short-term phase stability.

Start-time synchronization refers to the problem of aligning the beginning of the simulation scenario with the initialization of the receiver. This can be an issue when attempting to match a truth or reference dataset to measurements taken from the receiver in order to assess receiver performance. Although the timing information produced by the receiver's PVT solution is regularly used for synchronization, it may not always be available and it may not always be advisable to rely on the receiver in calibrating a test setup. It is also an issue when the receiver is excited by more than the GNSS simulator alone. For example, the simulation scenario may consist of a combination of simulated GNSS signals, other RF signals such as interferences, and/or simulated non-RF sensors such as inertial measurement units, magnetometers, barometers, wheel-speed sensors, etc. In this case, care must be taken to ensure that different sets of simulated data are precisely aligned.

Start-time synchronization also affects the repeatability of certain experiments. For example, it will vary the entry point into the received navigation message which, in turn, will affect receiver performance metrics such as the time to first fix, that is, the time needed for the receiver to compute and output its first PVT solution fix. In practice, such effects are averaged out in a well-executed experiment (it is common to

compute *mean-time to first fix*); nonetheless, for certain tests, a high degree of repeatability can be very useful. Many simulators provide a programmable *trigger* output or input to help tackle these issues. While this generally proves a satisfactory solution, it does require that the receiver can accept/produce a trigger input/output and that the user has the means to use it effectively.

Calibration is another significant issue for RF simulators, including an initial factory calibration, periodic maintenance, and experimental setup. In general, the precision of test equipment must be at least an order of magnitude greater than that of the device under test. For high-precision testing, simulation accuracy at the millimeter level for absolute and relative signal delay and less than 0.1 dB for signal power may be required. In such cases, the simulator may require careful calibration, in particular when more than one hardware channel is used. Many simulator manufacturers will correspondingly recommend and/or provide periodic maintenance for their products.

As mentioned previously, RF simulators are particularly useful for conducting coexistence or interoperability testing between GNSS signals or systems. When performing such tests, a user may be interested in

assessing the receiver performance when it is processing: one or more GNSS systems simultaneously; one GNSS system in the presence of other GNSS systems broadcasting at the same frequency; or, one or more GNSS systems in the presence of non-GNSS systems, such as, for example, pseudolite or satellite telephony systems. Signals from these non-GNSS systems or interferences can be directly combined with the simulated RF signal ensuring that the relative power levels are suitably adjusted. The ease with which various signals can be combined with the simulated RF signal for testing purposes highlights one particular vulnerability of the technique: a user can never be completely sure that the receiver is not also receiving and processing other interferences (that are not intended to be part of the test plan) from the laboratory. In particular, in a well-stocked laboratory, it is not uncommon to have many other devices capable of generating signals in the GNSS band, for example, signal generators, arbitrary waveform generators, or programmable RF transceivers. Even nearby devices which operate some hundreds of MHz away from the GNSS bands can emit spurious signals or harmonics which can find their way into the receiver-under-test RF feed which can lead to unexpected results.

18.3 IF-Level Simulators

IF GNSS simulators share many commonalities with the software-defined RF simulators discussed in Sect. 18.2.1 – the obvious difference being the absence of DAC, and up-conversion to RF – but are a more specialized type of simulator and far less widely used than their RF counterparts. Nonetheless, in certain applications, typically research and development, they have many distinct advantages. IF simulators are invariably implemented in software (or another easily reconfigurable platform) and produce one or more streams of digitized IF samples, similar to what a receiver would produce at the output of a digitizing front-end. The simulator must not only model the signal as it appears at the output of the receiver antenna, but it must also model the receiver RF and IF chains including the effects of the local oscillator, filtering and digitization; this is the other key difference from software-based RF simulators.

Applications of IF simulators include testing of software-defined receivers, testing of various parts of receiver hardware, or cases where it is desirable to prototype an algorithm or system in software, and/or is necessary to have an absolutely repeatable test scenario. Generally speaking, IF simulation cannot be used for

standard commercial-off-the-shelf receivers as their use necessitates access to an intermediate stage of the receiver and, for this reason, they are most commonly, if not solely, used for research, development, and educational purposes. As a result, they are most commonly encountered as in-house research tools, rather than as commercial products. The exception is the special case where the simulated IF signal is used in conjunction with other hardware, capable of up-converting and rebroadcasting at RF as with products from, for example, Avera, National Instruments, M3 Systems, and RACELOGIC. These record and playback systems are becoming increasingly popular and are discussed in Sect. 18.4.

One key benefit of IF simulators is flexibility. In general, the development cycle for software is much faster than that of hardware, so IF simulators are typically the easiest and fastest to be adapted to embody new signals, scenarios, and features than hardware-bound RF simulators (this is also a key motivator for the move toward software-based RF simulators discussed in Sect. 18.2.1). Moreover, the output of an IF simulator is itself quite easily manipulated by a user, thus allowing immense freedom to explore receiver performance.

This flexibility also allows a user to explore beyond the limitations of current technology, for example, synthesizing signals at arbitrary sample rates and/or arbitrary digitizer resolution.

Another important benefit of IF simulation is that the most basic and simple models for received GNSS signals are those of perfect signals. Indeed, all nonidealities must be consciously synthesized and applied to the synthetic signal. They can, therefore, be included or excluded at will and are absolutely repeatable, including noise. This enables a user to precisely investigate the impact of various signal and front-end features with relative ease, by including or excluding these features. Features such as receiver filters, reference oscillators, digitizers, gain-control algorithms, interference, atmospheric effects, and so on, can be excluded, included, and modified from experiment to experiment, while retaining the absolute repeatability of all other factors.

Being an entirely discrete-time simulated signal, when applied directly to the digital signal-processing stages of the receiver under test, IF simulators have the distinct advantage that they can operate faster than real time without sacrificing fidelity. This can happen because the IF simulator and receiver's digital processor need not wait for a full sample period to elapse, as is necessary in real operating conditions. Instead, with sufficient processing power, the IF signal samples can be generated and processed faster than if the signal were being sampled in real time (i.e., less than the sample period). For example, a sample stream that has been generated at a synthetic rate of 10 MHz might be processed at a rate of some tens or hundreds of megahertz. This capability can be an important consideration when a large range of tests must be conducted.

18.3.1 Implementation

If the IF samples are not to be upconverted to RF for re-broadcast, the IF simulator is required to model more than its RF counterpart as the effects of the receiver RF and IF stages must be included. Dominant amongst these effects are thermal noise in the amplification stages, the RF and IF filtering, the local reference clock, and frequency synthesizer, any on-board interference that may be introduced, the automatic gain-control unit, sampling, and digitization [18.21]. An IF simulator will generally create a baseband representation of the necessary signal modulation and, subsequently modulate it onto the appropriate IF, implement the appropriate filtering or conditioning, down-sample if necessary, and finally digitize [18.22].

The order in which these various effects are introduced into the signal throughout its generation is important as many of the effects are nonlinear. Non-

linear effects such as sampling, digitization, and gain-compression not only influence the amplitude of the signal, but also influence their spectral characteristics. For example, filtering an amplified signal, and amplifying a filtered signal, will produce different results if there are nonlinear elements in the amplifier [18.23]. In particular, the ordering of the sampling and digitization models is critical [18.22]. Of course, being a digital system, signals produced by an IF simulator are already discrete in time and so special attention must be paid to accurately modeling sampling effects. Under normal operating conditions, these effects would be negligible. However, in the case that the simulation must model some jamming or interference scenarios in which high power levels are present, these effects can become prominent.

In the case that the sampling model must exhibit/model some aliasing effects, it is likely that the IF data is synthesized at a much higher sample rate than is finally produced, such that intermediate filtering stages and the effects of aliasing can be appropriately incorporated. To consider a hypothetical example, if an IF dataset were to be created at a sample rate of 5 MHz complex, it may first be synthesized at upward of 20 MHz, filtered, digitized, and then down-sampled, such that the appropriate spectral folding would be more accurately modeled. In many cases, however, this level of fidelity may not be necessary, and direct synthesis at the final sample rate may be adequate.

Similarly, when synthesizing signal amplitude and phase, although the final IF signal may only exhibit 1 bit or 2 bit resolution, it will have been first synthesized with much higher precision, possibly even with floating-point precision. This can also be important when maintaining phase and time accumulators for various simulated parameters where the dynamic range of variables can be a limiting factor.

A basic flowchart of an IF signal simulator is shown in Fig. 18.5. In such simulators, the entire process is discrete time, although it is not uncommon that different components of the simulation operate at different rates. Firstly, the synthetic time process must be generated, whereby the transmit time corresponding to each sample epoch must be calculated – refer to (18.4). When simulating a perfect clock, these epochs are spaced evenly and according to the chosen sample rate, however, when simulating an unstable clock, these time increments will vary. When modeling a clock that is running slow, these increments will be larger, and when modeling a clock that is running fast, the increments will be smaller. In essence, this module must produce the inverse of the time process of an unstable clock or, more specifically, the projection of even time increments through this function [18.24–26].

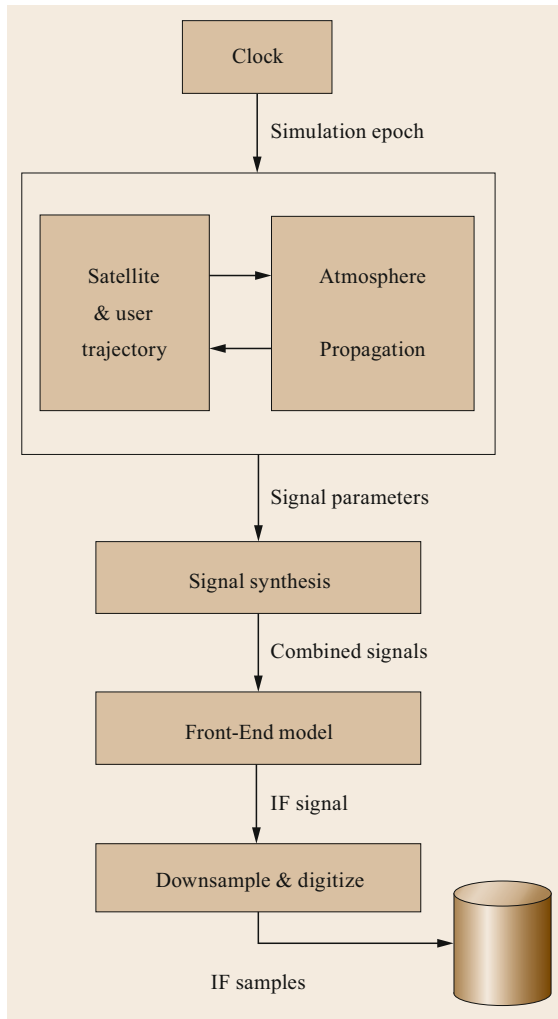


Fig. 18.5 Flowchart for an IF signal simulator

For each sample epoch, the physical model must be resolved, including transmitter and receiver trajectories and the propagation channel in between. In comparison to the transmitted signals, this model is slow moving and is typically updated at a much lower rate, and interpolated as necessary, for example, at a rate of 1 kHz, or even as low as 50 Hz. Having resolved this process, a set of signal parameters including amplitude and code- and carrier-phases for the LOS and multipath signals and other interferences are generated, representing the signals as they arrive at the receive antenna. These signal parameters are then used to synthesize *infinite*-resolution, noise-free, discrete-time signals which are combined into one sample. In the case that the simulator is implementing multiple front-end channels, these signals may be grouped accordingly.

This combined signal sample is then applied to an equivalent IF front-end model wherein the thermal noise floor is applied. This model may include one or more filtering stages, interspersed with nonlinear receiver elements such as amplifier models or active filter models. Also, depending on the fidelity of the simulation, this stage may include a down-conversion from high IF to low IF. The final stage of the simulation is gain control, down-sampling, and digitization. If the simulator has produced a high sample rate model to better model some of the previous stages, then down-sampling to the final sample rate is conducted here. The signal is then scaled to reflect automatic gain control and quantized to the required resolution.

18.3.2 Important Considerations

Software IF simulators operate in a purely deterministic manner. This feature can, in some cases, be a distinct advantage; however, in other cases, this can pose some difficulties which must be handled with care.

In terms of simulation repeatability, scenarios can, if desired, be absolutely repeatable which can be a great enabler for empirical or brute-force tuning of systems or receiver comparison. Of course, if absolute repeatability is not desired, deliberate steps must be taken to vary or randomize the simulation. At first glance, this may seem like a trivial task, but in reality, producing high-fidelity random processes can be very difficult when a large volume of simulation is necessary [18.27].

Synthesizing stochastic processes in software can be very difficult and requires a number of key ingredients: a reliable source of random (or pseudo-random) data; an appropriate model for the stochastic process; and a deterministic algorithm which can impress this model upon the random data. If any one of these ingredients is missing, the stochastic process cannot be faithfully replicated. The first ingredient, a source of random data, is readily available on almost all software platforms in the form of pseudo-random number generators that generally provide uncorrelated data exhibiting a bounded uniform distribution. For example, the current International Organization for Standardization (ISO) C standard implements a linear congruential generator for its random function [18.28]. In reality, many simulation scenarios require more high-quality pseudo-randomness and implement a dedicated generator [18.27, 29]. As such, it is generally one of the two remaining ingredients, either the model or the deterministic algorithm, that limits the simulator.

The fidelity of synthesized data is directly related to the accuracy of the model from which it is derived. In many cases, the models available for real-word phe-

phenomena are poor, either because the volume of data required to represent the phenomena is too large, the phenomena is too difficult to accurately measure, or because it is simply not cost-effective to launch a suitable data-collection campaign. In many cases, simulators either implement the generally accepted model in the field, or offer the user a choice of some of the more popular models. Indeed, even current simulators struggle to accurately model phenomena such as ionospheric scintillation due, primarily, to a lack of data upon which to base a simulation model.

The final limiting factor is the availability of a suitable deterministic algorithm which can convert the random data into the appropriate random process. This algorithm generally alters the distribution and the temporal correlation of the random data. In many cases, this involves the production of Gaussian variables that

have a well-defined temporal correlation, for which simple linear algorithms exist [18.25, 30]. Other processes which occur widely in nature are not so easily dealt with; oscillator instability being a classic example, wherein the stochastic phase process exhibits, so-called, fractional or *flicker* noise, which is exceedingly difficult to reproduce in software [18.25, 31–33].

Finally, we note that IF simulation cannot be used to perform a full receiver test, as the front-end is omitted from the test path. To include the front-end the IF data must be up-converted to an analog RF feed and rebroadcast. It is worth mentioning also that IF simulation need not necessarily be restricted to purely synthetic data, as genuine GNSS signals that have been pre-recorded by a suitable front-end device can be applied to the receiver under test. These two topics will be discussed in the following section.

18.4 Record and Playback Systems

Record and playback systems have increased in popularity in recent years as an interesting alternative or compliment to both RF and IF simulation techniques. Record and playback systems consist of two tools: a recording tool which captures one or more specific RF bands that are down-converted to IF prior to sampling, digitization, and storage to a disk; and a playback tool which up-converts these samples to an analog RF signal for rebroadcast or direct feed to the receiver (Fig. 18.2). Used either in isolation or in combination, these tools offer a wide range of possibilities to a user.

Record and playback systems overcome many of the challenges faced by traditional simulation techniques. For example, they remove any limit to the number of satellite signals, interferences, multipath propagation effects that can be recorded, as well as the need to model real-world effects such as the atmosphere or propagation channels. Indeed, the fidelity of the signal is only limited by the quality of the unit itself.

Record units can be used to capture rare phenomena, such as ionospheric activity, or difficult to (re)produce scenarios, such as complicated interference/spoofing scenarios or fast-fading propagation channels. In general, the technique is useful whenever it is infeasible to synthesize the scenario either due to its rarity, sparsity, geographic isolation, chaotic nature, or because it is simply not fully understood.

A user may even blindly record the appropriate RF spectrum and, therefore, subject the receiver to a realistic scenario, even if the user is not fully aware of the

scenario. Indeed, record units are not even restricted to capturing GNSS signals, and many record and playback systems facilitate the capture one or more RF bands, in the range of kilohertz to a few gigahertz.

Similarly, a user does not need to bring all of the devices on the test campaign; rather, the data can be played back at a later date to one or more devices. As a result, the use of record and playback systems has resulted in many laboratories compiling libraries of useful and/or interesting test scenarios, for future testing or research. These scenarios can then be distributed to third parties, for example, as in [18.16].

In cases where this data is used for performance assessment or calibration, and a truth or reference trajectory is required, record units are often used in combination with a GNSS reference receiver and/or high-quality inertial measurement units.

As mentioned previously, playback units are versatile and have been applied to the rebroadcast of IF samples collected from live GNSS signals, those collected from hardware RF simulators and, finally, to broadcast synthetic IF samples. As such, the combination of an IF simulator and a versatile replay unit constitutes a very practical and flexible RF simulator unit. In fact, this is the *modus operandi* of many modern RF simulators, as discussed in Sect. 18.2. Other examples of record and playback systems include Avera's RP-3200 Wideband RF Record & Playback system, RACELOGIC's LabSat 3 GPS Simulator, and Spirent's GSS6400.

18.4.1 Implementation

Unlike GNSS simulators, which are purpose-built and specialized tools, record and playback systems are often general-purpose devices. Historically, GNSS simulators have developed alongside GNSS receivers and have added or adapted functionality as testing demands and requirements necessitated. In contrast, record and playback systems, having their origins in the telecommunication industry, have been developed with a large and diverse user base. As a result, they offer a wealth of possibilities to a GNSS user, including the flexibility to capture far more than the GNSS band alone, and a digitization resolution far in excess of what is typically employed in GNSS systems; USRP-based systems are good examples of this, typically offering megahertz to gigahertz tuning range with 8 or 16 bit digitizer resolution.

Owing to their applications outside of GNSS, many record and playback systems are modular and can be manually configured to operate at various frequencies, typically between hundreds of kilohertz and 2–5 GHz. A typical unit will offer digitizer resolution ranging from 8 to 16 bit and will offer sample rates ranging from the low megahertz to some tens of megahertz. Units may vary in the number of channels they offer and, depending on the architecture, these channels may be either sample- or phase synchronized. For more portable solutions, however, many researchers opt for a GNSS dedicated solution, having a more restricted radio band and, typically, lower resolution digitizers.

Example record and playback systems are shown in Figs. 18.6 and 18.7, respectively. Both units are from National Instruments but differ greatly in terms of size and capability. The USRP units shown in Fig. 18.6 are highly portable but each is limited to recording a single band, and no more than two can be (synchronously) combined together. Furthermore, the data throughput is the same regardless of whether one or two units are recording data. The USRP unit can also be used to replay the data they recorded.

In contrast, the system shown in Fig. 18.7 is much larger and each chassis can be configured to log from multiple bands simultaneously, and multiple chassis can be time synchronized to collect even more data. Use of an array of hard drives allows for very high data recording rates across all bands (up to 40 Mb/s each).

The record stream generally consists of an antenna, amplification and down-conversion stages, digitization, and an optional signal conditioning and down-sampling stage. This digital data stream is either stored within the unit or streamed to a host computer. The replay channel essentially consists of the reverse: digital to analog, up-conversion in one or more stages and either power

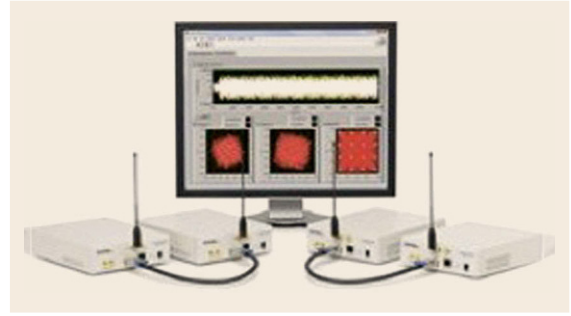


Fig. 18.6 Four National Instruments USRP units, each capable of logging data from a particular band. Pairs of units can be time synchronized using special cables (*lower-left* and *right*) in order to log two bands at once (courtesy of National Instruments)

amplification or attenuation, depending on whether the signal is broadcast or fed directly to a receiver.

In most cases, the system operates in half-duplex mode, where data is first recorded and later rebroadcast, which generally meets the demands of receiver testing. Full-duplex mode, where the unit simultaneously receives and transmits, is offered by some units [18.34], and can facilitate some more complicated tests. A full-duplex system is capable of, for example: performing loop-back, where the received signal is simply rebroadcast as is; first modifying the received signal, as digital IF, prior to rebroadcasting; or creating a simulated signal based on the content of the received signal. These concepts are investigated further in Sect. 18.6.

18.4.2 Important Considerations

When using a record and playback system, there are a number of important factors that should be considered when designing an experimental setup.

Gain control and dynamic range are important factors. Firstly, it is important that the record unit has a significantly higher dynamic range than the receiver, both in the ADC and DAC stages, and, secondly, the converter gains should be held constant during the record and playback stages. Automatic gain control in the record unit is not advisable for receiver testing as, although the transmitted signal-to-noise ratios may remain constant, the receiver under test may perceive the variations in signal and noise floor separately, resulting in the appearance of signal-to-noise-ratio variations.

Ideally, the bandwidth of the record and playback system is greater than the bandwidth of the receiver under test, such that the playback system does not induce any ripple or roll-off the portion of the transmitted signal that lies within the receiver's passband. Further-



Fig. 18.7 A record and playback system installed at the Joint Research Center, Italy, comprising of two national instruments chassis, one housing a vector signal transceiver and the other housing pairs of up- and down-converters and digitizers and analog-to-digital converters (ADCs). A single-reference oscillator is used to clock both chassis, and a calibrated set of trigger signals are generated to simultaneously launch record and playback. IF samples collected in four configurable bands are streamed to and from an array of magnetic hard drives (*bottom*)

more, when operating a record and playback system, there is no possibility to replay the GNSS signal alone, as it has been captured with thermal noise present. In this case, the noise figure observed in any subsequent testing has been limited by the noise figure of the record device. The properties of this noise are quite important. When thermal noise is simulated, it must appear white to the receiver, specifically, it should exhibit the same shape as the frequency response of an appropriate antenna which, from the perspective of most receivers, is flat. For simulators, this is generally not a problem. Record and playback systems, in contrast, may be bandwidth limited. Commercially available record and playback systems have configurable bandwidths ranging from 1 to 100 MHz. When performing a playback test using a broadcast bandwidth that is noticeably narrower than the receive bandwidth of the receiver under test, some problems may arise, for example, within the

receiver automatic gain control, interference mitigation algorithms, or even in tracking.

One simple method of circumventing the problems of low playback bandwidth is to combine it with live thermal noise. The rebroadcast data can be passed through a variable attenuator to bring the playback spectrum close to the thermal noise floor. Subsequently, it can be passed through a broadband power amplifier to bring the narrowband playback signal plus thermal noise up to the power level that the receiver under test is expecting, similar to that of an active antenna.

Once again, the quality of the oscillator used is very important; however, unlike RF or IF simulator systems, frequency accuracy is less important as frequency offsets introduced in the record process will cancel in the up-conversion process, provided the oscillator is reasonably stable between the record and playback times. Short-term stability, on the other hand,

is amplified. White, flicker, and random-walk phase and frequency noise will be uncorrelated between record and playback times and so their effects will combine and manifest themselves as apparent errors in the *receiver* clock. It is important, therefore, to ensure that the reference clock used for the system is an order of magnitude more stable than that of the receiver under test.

Although record and playback systems do not impose a limit on the number of GNSS signals that can be reproduced, unlike traditional simulators, data transfer can impose limitations on the total bandwidth captured. Recording IF data can produce quite a lot of raw data that cannot be effectively compressed, as it is generally dominated by thermal noise. The task of transferring and storing this data is generally the limiting factor for these systems. The total data rate will be a linear function of the ADC/DAC resolution, the sample rate, and the number of frequencies recorded. For example, a single-frequency system sampling data at 20 MHz complex with 8 bit resolution per sample generates 40 Mb of data per second, or 2.4 Gb/min. Higher resolution, multifrequency systems will produce proportionally more. Even when the necessary recording

rates are attained, tests with record and playback systems are generally limited to a few tens of minutes or a few hours for practical reasons, which has implications for experiment design.

Another factor that influences experiment design when using a record and playback system is the availability of accurate truth data. While with traditional simulators, a full set of truth data is available and receiver performance can be assessed in an absolute sense, record and playback systems offer no such feature. A user must either resign to assessing performance relative to a reference receiver, or to sourcing reference data elsewhere, for example, via precise ephemerides and a known trajectory or by including other sensors such as an inertial measurement unit (IMU). In either case, there is a limit to the degree of accuracy to which the truth data can be estimated because many variables, such as atmospheric and channel propagation parameters, are unknown. A user must therefore take care not to attempt to measure performance with greater precision than that of the truth data. Ideally, experiments conducted via record and playback should be designed to be as insensitive as possible to the accuracy of the truth data.

18.5 Measurement-Level Simulators

Measurement-level simulation aims at testing navigation solution calculations that are implemented internal to the receiver's firmware or external to the receiver in the form of third-party software. This differs from previously discussed simulators in three key ways.

First, whereas RF simulators generate a continuous signal feed and IF simulators need to generate millions of samples per second of simulation, measurement simulators are typically concerned with inputs on the order of 50 Hz or less. This is driven by the fact that, with a few exceptions, receivers simply do not need such high data rates. Correspondingly, allocating more processing resources to generate measurements at a higher rate is unjustified. This makes measurement simulation much less computationally strenuous.

Second, as mentioned above, the signal-processing stage of a receiver is insensitive to the low-frequency errors caused by orbital errors and the atmosphere, for example, and pass through this stage unaffected. These errors are therefore present in the pseudorange, carrier-phase, and carrier Doppler measurements, and need to be modeled with the level of fidelity needed for the type of testing being conducted. As with all simulations, the level of fidelity will vary greatly between applications.

The third difference is that measurement-level testing combines outputs from different channels within the receiver. Testing with RF and IF simulators is often focused on evaluating the signal-processing stages of a receiver and is thus concerned with channel-level information such as acquisition and tracking thresholds, C/N_0 estimation, time to first fix, etc. In contrast, testing the navigation solution combines information from all channels together. By extension, the fidelity of a single satellite's signal(s) is no longer sufficient and care must be exercised to ensure that measurements to different satellites adequately reflect the relative errors that are expected. The most obvious source of error variation arises from tracking satellites at different azimuth and elevation angles which induces different errors due to the atmosphere, but multipath errors are also critical in this context.

A fourth difference is also applicable in situations where differential processing between two or more receivers is being tested. In these cases, in addition to modeling the relative errors between satellites at a single location, the *spatial* variability of the errors is critical, as this ultimately dictates the accuracy of such systems. For carrier-phase processing in particular, the spatial variability of the errors is most important since

this will have the most profound impact on the ability to resolve the integer ambiguities.

In light of the above, measurement-level simulators are typically used as a means of testing a navigation solution's sensitivity to the magnitude and/or variability (temporal and/or spatial) of certain errors and, by extension, its ability to model or otherwise account for these errors [18.35, 36]. For applications where multipath is the largest effect (meaning the error cannot be practically modeled), the ability of the navigation solution to identify and reject large errors using fault detection and exclusion (FDE) is equally important, if not more so.

Measurement simulators are most commonly used for in-house development and testing of receiver firmware and/or separate data-processing software. They also play an important role in research and development and are thus commonly found in academia and research institutes.

Another motivation for using measurement simulators is to assess PVT performance as a function of the number of GNSS, the number and type of signals, and the number of frequencies used as input to the navigation solution. In turn, this can be used to decide how receiver resources are allocated for a particular application.

18.5.1 Implementation

In many instances, measurement simulation is *almost* a natural by-product of RF or IF simulators. Since the measurement errors need to be known in advance (18.3), they are usually logged to file by the simulator. Indeed, these measurements form the basis of a truth or reference dataset for RF or IF simulation. These files can then be read and used to form observations absent of noise and multipath effects.

Another form of measurement-level simulation can be found in the analysis of a GNSS constellation. For example, the early stages of the Galileo constellation design were assessed using service volume simulation, to reproduce the functional and performance behavior of the Galileo system. Generally used to conduct navigation and integrity performance analyses over long time periods and over large geographical areas, this tool computes figures of merit such as visibility, accuracy, integrity, continuity, and availability of service. However, it is also capable of generating raw Galileo and GPS observables for the validation of navigation and integrity processing facilities [18.6, 37].

Regardless of the implementation, the key challenge is to generate errors that exhibit the range of magnitudes, temporal variability, and spatial variability consistent with reality. This challenge is not to be taken lightly. Although many different models exist to correct

for ionosphere and troposphere errors (Chaps. 38 and 39), these models are not perfect, meaning that some residual error remains in the measurements after the models are applied. Typically, models tend to produce overly smooth errors, both temporally and spatially. Using such models for simulation thus leads to overly optimistic performance because the simulator and receiver are using the same, or very similar models, thus limiting the amount of testing that can be reliably performed.

Inclusion of the effect of noise and multipath typically requires complete knowledge of how a particular receiver operates including: the length of coherent and/or noncoherent integration performed; the correlator spacing, and the types, order, and bandwidths of the tracking loops, etc. [18.38]. In many cases, these parameters are unknown to the user. Moreover, in some cases – when developing software to work with a variety of receivers, for example – specific parameters are less useful than covering a wide range of possible parameters. In light of this, noise and multipath models are usually based on some form of random process whose general statistical characteristics approximate those empirically obtained by a particular receiver in a particular environment as, for example, in [18.35, 39, 40].

In the case of benign environments (e.g., open sky), random noise can be added to match what is normally obtained for a particular receiver, usually scaled by the simulated C/N_0 of the signal. However, for high-sensitivity receivers, the effect of shadowing should be included, as per (18.9). In addition, the effect of fading also needs to be included explicitly.

Multipath errors can also be generated using a statistical model. A common approach is to model the errors as Gauss–Markov processes [18.40–43]. Alternatively, if a particular reflection geometry is assumed, the instantaneous effect of multipath on the code and carrier-phase discriminators may be used. Note, however, that this must assume the receiver response to the multipath signal, including filtering effects provided by the antenna and tracking loops, and any receiver-based multipath mitigation schemes.

18.5.2 Important Considerations

There are two main considerations when using measurement simulators. First, as mentioned above, proper simulation of noise and multipath normally requires the knowledge of how a particular receiver is implemented. If these parameters are unavailable, then assumptions are made based on empirical evidence.

The second consideration is how the type of processing undertaken compares with the errors being

simulated. In general, pseudorange processing is well suited to measurement simulation because noise and multipath are the dominant errors meaning that the fidelity of the ionosphere and troposphere errors is less of a concern. This is even more true for differential processing where the effect of ionosphere and troposphere is virtually zero (relative to pseudorange noise and multipath), except for over extremely long receiver separations.

Extra care must be exercised when using measurement simulation for high-sensitivity receivers. Whereas *standard* sensitivity receivers do a reasonably good job of only producing measurements containing errors within an acceptable range, high-sensitivity receivers may occasionally produce highly erroneous measure-

ments. This occurs when the receiver is working at or near its tracking threshold. Correspondingly, accurately simulating a high-sensitivity receiver at the measurement level should also account for these *threshold effects*.

In contrast to pseudorange positioning, differential carrier-phase positioning algorithms are much more sensitive to the fidelity of the ionosphere and troposphere errors. The reason is that the reduced multipath errors on the carrier-phase means that these other errors can be significant – or even dominate – as the receiver separation increases. If the models are spatially and/or temporally too smooth, this could have significant impacts (positive or negative) on the ability to reliably resolve the integer carrier-phase ambiguities.

18.6 Combining Live and Simulated Data

One use of simulators that is noteworthy is the combination of live and simulated stimuli. Although this approach does not constitute a traditional simulator, and the style of testing is less well defined, it is becoming increasingly popular due, in part, to: the increasing number of GNSS satellites in view of a typical receiver; the increasing number of frequencies on which satellites transmit; the increasing use of other sensors in GNSS receivers; and the increasingly complicated scenarios a user wishes to examine.

A simulator tool may be limited in the number of signals/sensors it can synthesize and/or the number frequencies it can simulate. Fortunately, it is often the case that a user is only interested either in controlling, or examining receiver performance relative to a small subset of the signals in view. In such a case, it is possible to provide the GNSS receiver under test with both genuine signals, of which the user may or may not have complete knowledge, combined with a set of simulated signals, over which the user has full control and for which the user has good truth data.

Another common reason to augment both simulated and live data is when the user has no readily available means of simulating a particular signal, but has the means to record it live. Combining of genuine and simulated signals can either be done live, where the simulated component is produced in real time and fed directly to the receiver under test; or can be conducted in post-mission, where the genuine data is stored and reproduced (for example, via record and playback system) for combining with the simulated signals.

In some cases, this form of augmentation is simple and requires little or no calibration or synchronization, such as jamming or interference scenarios. In other

cases, synchronization is critical, for example, when simulated GNSS signals are combined with live ones.

Augmentation of live GNSS signals with simulated signals or augmentation of a simulation with live signals is commonly employed to overcome either limited simulation resources or unavailability of live data including, for example:

- In the case that a user wishes to examine the performance of a multi-GNSS receiver on a GNSS system which is not yet fully deployed, for example, Galileo or BeiDou. A user may simulate only the satellites that are missing, providing the receiver with genuine signals for the remainder. This may either be done in real time or in post-mission via record and playback of the genuine signals. In practice, however, the post-mission case is far more straightforward.
- When an accurate model or simulation resource of non-GNSS signals is not readily available, for example, for compatibility/vulnerability analysis with devices such as jammers and pseudolites. In these cases, it may be necessary to have a repeatable GNSS component with an accurate truth dataset for performance evaluation. However, it may not be necessary to know exactly what the properties of the non-GNSS component are. Thus, the GNSS component might be simulated, while the non-GNSS component may be broadcast/conducted live [18.16, 17].
- In the case that a user wishes to add channel effects of the genuine received satellite signals to the RF feed of the receiver. A user may wish to add simulated multipath reflections to a scenario, in which

case the LOS signals are genuine and the simulator is used to add the reflections in a controlled and deterministic fashion.

- Similar to the multipath case, if a user wishes to simulate a spoofing attack. In this case, the simulator is used to produce the spoofed signals which are then either combined with the genuine LOS signals at the RF feed directly, or broadcast toward the target receiver from a local antenna.
- In the case when GNSS signals and other sensor data needs to be combined. In such a case, sensor data may be collected along some well-known reference trajectory. The GNSS signals corresponding to this reference trajectory can be subsequently simulated and combined with the sensor data. This might allow a user to examine integrated performance. Moreover, the quality and availability of the GNSS signals are readily controllable by the user. One example of this might be the combination of genuine inertial measurements with GNSS measurements for a tightly- or loosely-coupled solution.

18.6.1 Implementation

As mentioned above, the augmentation of simulated and live signals is conducted under exceptional circumstances. The requirements in such circumstances can vary widely, and this style of testing is conducted using a range of independent modules, rather than one bespoke unit.

Elements that may be employed include those required for simulation; sufficient equipment to observe or capture the genuine signals; and, where necessary, playback elements. Apart from these obvious elements, combining signals from different sources may also require some specialized tools for configuration and setup, including network/spectrum analyzers, timing, or synchronization tools and a reference receiver.

18.6.2 Important Considerations

As with pure simulation and record and playback methods, the effectiveness of augmenting simulation with genuine signals is limited by the quality of the combined signals. In this regard, the relative timing and relative power levels are amongst the most important. To make any high-precision measurements on the receiver under test, it is necessary that a user can distinguish between errors introduced by the user in the combining of the various signals, and those which relate to receiver performance. Moreover, it is important

that the combined signals accurately reflect the desired test scenario.

For example, when combining simulated and genuine signals at RF, it may be beneficial to employ a network and/or spectrum analyzer to ensure appropriate matching of the network. In particular, combining live and simulated signals in a laboratory can necessitate long cabling and can be a significant source of error. Care must be taken to limit power loss, minimize interference emission, and signal reflection.

When injecting strong interference signals at RF, the reverse isolation of the network should be considered, both to ensure the quality of the test and to avoid unintentional broadcast. Matching power levels can also be a difficult task as the genuine signals and thermal noise will have been amplified at least once prior to combining. In general, ensuring that the simulated signals conform to the genuine signals involves scaling the power of the simulated signals relative to the thermal noise floor observed at the genuine signal feed.

When genuine GNSS data is combined with simulated GNSS data, timing becomes a crucial consideration. The passage of time must be consistent between the simulated and genuine data, and therefore biases in the reference oscillator used to drive the simulator must be bounded and the delay induced by the combining network must be accounted for in the generation of the simulated signals.

Firstly, the passage of time in the simulator must be reasonably consistent with the true passage of time. Unlike pure simulation or record and playback scenarios, where the simulator or the playback system *defines* system time, this style of augmentation necessitates consistency. To achieve this, the system should be equipped with either a GNSS-disciplined reference clock, or a complete reference receiver, which can serve to synchronize or discipline the simulator.

Secondly, the system time used by the simulator must lead the true system time by the difference in electrical length between the antenna–receiver path and the simulator–receiver path. This calibration is particularly important when the simulator is used to generate signals from a GNSS system that is also present in the set of genuine signals, for example, when performing multipath or spoofing testing. It is less crucial when simulating signals from a system not present in the set of genuine signals as this calibration error will simply appear as an intersystem bias.

Overall, this style of simulation can be very powerful and is an enabler for research and development; however, users should be cautioned that it is highly sensitive to calibration.

18.7 Other Considerations

The GNSS simulation and testing landscape is rapidly evolving to keep pace with the evolution and requirements of GNSS receiver manufacturers and GNSS system integrators. Correspondingly, there are many different features currently available, and more will likely become available in the future. This section briefly presents some considerations for deciding what type of simulator to use as well as what features might be of interest for a given type of simulator.

18.7.1 GNSS Systems Supported

The most obvious considerations revolve around the generation of the GNSS data including the GNSS systems that can be simulated, how many frequencies for each system, and how many signals/satellites can be simulated at a time. Also of importance might be the ability to simulate satellite- or ground-based augmentation system (SBAS or GBAS) signals. The number of outputs (e.g., RF feeds or IF sequences) that can be generated at a time is important for differential processing and testing of attitude determination systems, for example.

18.7.2 Interference and Spoofing

Both military and civilian systems are concerned with the effect of interference/jamming and spoofing (Chap. 16). Proper testing of a receiver in their presence requires the same level of control as testing a receiver in their absence. To this end, the types of interference that can be simulated (e.g., continuous wave, swept wave, pulsed, broadband, etc.) and their dynamic range are critical parameters that should be considered. Similarly, the types of spoofing scenarios that can be generated should also be taken into account.

However, interference/jamming need not be intentional and may arise from other RF signals/systems operating in adjacent frequency bands, either through insufficient filtering at the transmitter or via inadvertently re-radiating of interfering signals within the device. Possible signals may include mobile phones (e.g., LTE, 3G, GSM, etc.), short-range communications (e.g., Bluetooth, WiFi, etc.), or analog television broadcast [18.44]. A simulator's ability to generate these signals should therefore be evaluated.

18.7.3 Other Data

As GNSS continues to be integrated with other sensors or systems, the ability to test the fully integrated system,

instead of just the GNSS receiver in isolation, becomes an important part of a testing program. Some simulators are already capable of generating additional outputs that can be used for navigation including inertial measurement unit (IMU) data, WiFi signals (for ranging and/or proximity sensing), and assistance data for testing aided GNSS receivers. As other sensors or systems become more commonly integrated with GNSS, the ability of a simulator to include them in the testing chain may be critical.

18.7.4 Configurability

The ease of configuring and running a simulation scenario can save considerable time and money and should be given requisite consideration. This may include the ability to control the simulator remotely or via scripts, thus allowing users to work from their own workstation and to perform testing 24 hours per day.

In terms of scenario setup, some simulators allow standard log files (e.g., National Marine Electronics Association (NMEA, see Annex A.1.1) or Receiver INdependent EXchange (RINEX, see Annex A.1.2) formats) recorded from a receiver in a real-world environment to be used as the basis of a simulation. In these cases, the logged data contains information about the user's position, what satellites were in view at every epoch and their measured power level. Such an approach saves time and arguably makes the simulation more realistic and this method of simulation implies that the user already holds a reference trajectory in a suitable format (Sect. 18.4.2).

Other simulators allow users to define a receiver's surroundings in terms of possible reflectors and then uses this to automatically determine what satellites are in view as well as when a particular satellite receives direct and/or reflected signals.

18.7.5 Expandability

All simulators have some software interface and processing capabilities and are, thus, theoretically, easily expandable. Similarly, adding additional hardware, for example, to simulate more signals is also generally possible. Nevertheless, understanding what changes/additions can be made after purchase should be an important consideration when purchasing a simulator or testing system.

18.8 Summary

This chapter has looked at using simulators for testing the performance of GNSS receivers and, to a lesser extent, complete positioning systems. With simulator options ranging from RF- to measurement-level or combinations thereof, system designers have a number of resources at their disposal that can be used for assessing various parts of a GNSS receiver or a GNSS-integrated system.

As the simulation and testing landscape continue to evolve, new features and capabilities will surely become available. Ultimately, however, proper selection of equipment and careful design of tests – taking into account key objectives – are still a key part of the testing process.

References

- 18.1 G. Seeber: *Satellite Geodesy: Foundations, Methods, and Applications* (Walter de Gruyter, Berlin 2003)
- 18.2 P. Axelrad, R.G. Brown: GPS navigation algorithms. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B. Parkinson, J.J. Spilker (AIAA, Washington 1996) pp. 409–433
- 18.3 Global Positioning Systems Directorate: Navstar GPS Space Segment/Navigation User Interfaces, Interface Specification, IS-GPS-200H, 24 Sep. 2013 (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo 2013)
- 18.4 M.C. Moreau, E.P. Davis, J.R. Carpenter, D. Kelbel, G.W. Davis, P. Axelrad: Results from the GPS flight experiment on the high earth orbit AMSAT OSCAR-40 spacecraft, Proc. ION GPS 2002, Portland OR 24–27 Sep. 2002 (ION, Virginia 2002) pp. 122–133
- 18.5 E. Kahr: Prospects of multiple Global Navigation Satellite system tracking for formation flying in highly elliptical earth orbits, Int. J. Space Sci. Eng. **1**(4), 432–447 (2013)
- 18.6 M.J. Unwin, R. De Vos Van Steenwijk, Y. Hashida, S. Kowaltschek, L. Nowak: GNSS at high altitudes – results from the SGR-GEO on GIOVE-A, 9th Int. ESA Conf. Guid. Navig. Control Syst., Porto Portugal 2–6 June 2014 (ESA, Noordwijk 2014)
- 18.7 S. Satyanarayana, D. Borio, G. Lachapelle: A composite model for indoor GNSS signals: Characterization, experimental validation and simulation, Navigation **59**(2), 77–92 (2012)
- 18.8 S. Satyanarayana: GNSS Channel Characterization and enhanced Weak Signal Processing, Ph.D. Thesis (Univ. of Calgary, Calgary 2011)
- 18.9 J.T. Wu, S.C. Wu, G.A. Hajj, W.I. Bertiger, S.M. Lichten: Effects of antenna orientation on GPS carrier-phase, Man. Geod. **18**, 91–98 (1993)
- 18.10 A.K. Tetewsky, F.E. Mullen: Carrier phase wrap-up induced by rotating GPS antennas, Proc. ION AM 1996, Cambridge MA 19–21 June 1996 (ION, Virginia 1996) pp. 21–28
- 18.11 M.C. Vigano, C. Gigandet, S. Vaccaro: Wideband, phase-stable antenna for navigation applications, 7th Eur. Conf. Antennas Propag. (EuCAP 2013), Gothenburg, Sweden 8–12 Apr. 2013, ed. by P.-S. Kildal (2013) pp. 2226–2229
- 18.12 M.E. Cannon, G. Lachapelle, M.C. Szarmes, J.M. Hebert, J. Keith, S. Jokerst: DGPS kinematic carrier phase signal simulation analysis for precise velocity and position determination, Navigation **44**(2), 231–245 (1997)
- 18.13 M. Irsigler, B. Eissfeller: PLL tracking performance in the presence of oscillator phase noise, GPS Solutions **5**(4), 45–57 (2002)
- 18.14 T.E. Humphreys, M.L. Psiaki, P.M. Kintner: Modeling the effects of ionospheric scintillation on GPS carrier phase tracking, IEEE Trans. Aerosp. Electron. Syst. **46**(4), 1624–1637 (2010)
- 18.15 F. Zimmermann, T. Haak, E. Steindl, S. Vardarajulu, O. Kalden, C. Hill: Generating Galileo raw data – Approach and application, Proc. Data Syst. Aerosp. (DASIA 2005), Edinburgh, Scotland 30 May–2 June 2005, ed. by L. Ouwehand (ESA, Noordwijk 2005), pp. 45.1–45.12
- 18.16 T. Humphreys, J. Bhatti, D. Shepard, K. Wesson: The Texas spoofing test battery: Toward a standard for evaluating GPS signal authentication techniques, Proc. ION GNSS 2012, Nashville, TN 17–21 Sep. 2012 (ION, Virginia 2012) pp. 3569–3583
- 18.17 I. Petrovski, T. Tsujii, J.M. Perre, B. Townsend, T. Ebinuma: GNSS Simulation: A user’s guide to the Galaxy, Inside GNSS **5**(7), 36–45 (2010)
- 18.18 D.W. Allan: The science of timekeeping, IEEE Trans. Instrum. Meas. **IM-36**(2), 646–654 (1987)
- 18.19 R.L. Filler: The acceleration sensitivity of quartz crystal oscillators: A review, IEEE Trans. Ultrason. Ferroelectr. Freq. Control **35**(3), 297–305 (1988)
- 18.20 R.L. Filler, J.R. Vig: Long-term aging of oscillators, IEEE Trans. Ultrason. Ferroelectr. Freq. Control **40**(4), 387–394 (1993)
- 18.21 C.J. Hegarty: Analytical model for GNSS receiver implementation losses, Navigation **58**(1), 29 (2011)
- 18.22 J.T. Curran, D. Borio, G. Lachapelle, C.C. Murphy: Reducing front-end bandwidth may improve digital GNSS receiver performance, IEEE Trans. Signal Process. **58**(4), 2399–2404 (2010)
- 18.23 R. Price: A useful theorem for nonlinear devices having gaussian inputs, IRE Trans. Inf. Theory **IT-4**, 69–72 (1958)
- 18.24 N.J. Kasdin: Discrete simulation of colored noise and stochastic processes and 1/f power law noise generation, Proc. IEEE **83**(5), 802–827 (1995)
- 18.25 J.A. Barnes: Simulation of oscillator noise, Proc. 38th Annu. Symp. Freq. Control, Philadelphia, PA 29 May–1 June 1984, ed. by J.R. Vig (1984) pp. 319–

- 326, doi:10.1109/FREQ.1984.200775
- 18.26 H. Meyr, G. Ascheid: *Synchronization in Digital Communication: Phase-, Frequency-Locked Loops, and Amplitude Control* (Wiley, New York 1990)
- 18.27 S.M. Ross: *Simulation*, 5th edn. (Academic Press, Amsterdam 2012)
- 18.28 Standard for Programming Language C++, ISO Standard ISO/IEC 14882:2014 (International Organization for Standardization, Geneva 2014)
- 18.29 RANDOM.ORG Randomness and Integrity Services <http://www.random.org/>
- 18.30 G.E.P. Box, M.E. Muller: A note on the generation of random normal deviates, *Ann. Math. Stat.* **29**(2), 610–611 (1958)
- 18.31 B.B. Mandelbrot: A fast fractional gaussian noise generator, *Water Resour. Res.* **7**(3), 543–553 (1971)
- 18.32 W.C. Lindsey, C.M. Chie: Theory of oscillator instability based upon structure functions, *Proc. IEEE* **64**(12), 1652–1666 (1976)
- 18.33 J.A. Barnes: Characterization of frequency stability, *IEEE Trans. Instrum. Meas.* **IM-20**(2), 105–120 (1971)
- 18.34 National Instruments: Vector Signal Transceiver <http://www.ni.com/vst/>
- 18.35 N. Luo: Precise Relative Positioning of Multiple Moving Platforms Using GPS Carrier Phase Observables, Ph.D. Thesis (University of Calgary, Calgary 2001)
- 18.36 O. Julien, M.E. Cannon, P. Alves, G. Lachapelle: Triple frequency ambiguity resolution using GPS/Galileo, *Eur. J. Navig.* **2**(2), 51–57 (2004)
- 18.37 ESA: Galileo System Simulation Facility, <https://www.gssf.info/>
- 18.38 P. W. Ward, J. W. Betz, C. J. Hegarty: Satellite signal acquisition, tracking, and data demodulation. In: *Understanding GPS: Principles and Applications*, ed. by E. D. Kaplan, C. J. Hegarty (Artech House, Norwood 2006) pp. 153–242
- 18.39 M. Dumville, W. Roberts, D. Lowe, B. Wales, P. Pettitt, S. Warner, C. Ferris: On the road under real-time signal denial, *GPS World* **24**(5), 40–44 (2013)
- 18.40 A. Jahn, H. Bischl, G. Heiss: Channel characterisation for spread spectrum satellite communications, *Proc. 4th Int Symp. Spread Spectr. Tech. Appl.*, Mainz, Germany 1996, Vol. 3, ed. by P.W. Baier (1996) pp. 1221–1226
- 18.41 K.R.L. Edwards: Site-Specific Point Positioning and GPS Code Multipath Parameterization and Prediction, Ph.D. Thesis (Ohio State University, Columbus 2011)
- 18.42 J.F. Raquet: Multiple reference GPS receiver multipath mitigation technique, *Proc. ION AM 1996*, Cambridge, MA 19–21 June 1996 (ION, Virginia 1996) pp. 681–690
- 18.43 K. O’Keefe, M.G. Petovello, G. Lachapelle, M.E. Cannon: Assessing probability of correct ambiguity resolution in the presence of time-correlated errors, *Navigation* **53**(4), 269–282 (2006)
- 18.44 M. Wildemeersch, E. Cano Pons, A. Rabbachin, J. Fortuny Guasch: Impact Study of Unintentional Interference on GNSS Receivers, JCR Report EUR 24742 EN (European Union, Luxembourg 2010)

GNSS Algorithms

Part D

Part D GNSS Algorithms and Models

19 Basic Observation Equations

André Hauschild, Wessling, Germany

20 Combinations of Observations

André Hauschild, Wessling, Germany

21 Positioning Model

Dennis Odijk, Leidschendam, The Netherlands

22 Least-Squares Estimation and Kalman Filtering

Sandra Verhagen, Delft, The Netherlands
Peter J.G. Teunissen, Perth, Australia

23 Carrier Phase Integer Ambiguity Resolution

Peter J.G. Teunissen, Perth, Australia

24 Batch and Recursive Model Validation

Peter J.G. Teunissen, Perth, Australia

Basic Observ

19. Basic Observation Equations

André Hauschild

This chapter introduces the fundamental observation equations for multiconstellation global navigation satellite systems (GNSSs). It starts with an introduction of the basic observation equations for pseudorange, carrier-phase, and Doppler measurements. In the remainder of the chapter, the parameters used in modeling the basic observation equations are discussed. The parameters covered in the discussion are relativistic effects, atmospheric delays, the carrier-phase wind-up effect, antenna phase-center offset and variation, pseudorange and carrier-phase biases, and finally multipath errors and receiver noise.

19.1	Observation Equations	561
19.1.1	Pseudorange Measurements	561
19.1.2	Carrier-Phase Measurements	563
19.1.3	Doppler Measurements	563
19.2	Relativistic Effects	564
19.3	Atmospheric Signal Delays	565
19.3.1	Ionosphere	566
19.3.2	Troposphere	568
19.4	Carrier-Phase Wind-Up	569
19.4.1	Wind-Up Effect for Radio Waves	569
19.4.2	GNSS Satellite Attitude Modeling	570
19.5	Antenna Phase-Center Offset and Variations	572
19.5.1	Overview	573
19.5.2	Calibration Techniques	574
19.5.3	Examples for Phase-Center Variations	575
19.6	Signal Biases	576
19.6.1	Pseudorange Biases	576
19.6.2	Carrier-Phase Biases	578
19.7	Receiver Noise and Multipath	578
19.7.1	Receiver Noise	578
19.7.2	Multipath Errors	579
	References	579

19.1 Observation Equations

This section develops generic basic observation equations for pseudorange, carrier-phase and Doppler measurements. All necessary modeling parameters will be identified and discussed in further detail in the following sections.

19.1.1 Pseudorange Measurements

For the generation of pseudorange observations, a GNSS receiver measures the apparent signal travel time of the signal from the navigation satellite to the user. The receiver's delay lock loop (DLL) generates a replica of the signal's code based on its internal frequency source and aligns it with the received signal. The necessary time shift is a measure of the appar-

ent transit time modulo the code chip length. It is then combined with the number of complete code chips, complete code repeats, and additional information from the satellite's navigation data to obtain the unambiguous apparent signal travel time or, if multiplied with the speed of light, the apparent range or pseudorange. As shown in the following, the receiver's measurements differ from the true signal travel time or true range, since they are affected by the receiver's and the satellite's clock offsets with respect to the GNSS system time as well as other errors and signal delays.

The arrival time t_A of the signal at the receiver depends on the time t_E , when the signal was emitted at the satellite, the signal travel time τ between the satellite s and the receiver r , the relativistic effect due to

space-time curvature $\delta t_{\text{stc}}^{\text{rel}}$, the ionospheric delay I , and the tropospheric delay T as follows

$$t_A = t_E + \tau(t_A) + \delta t_{\text{stc}}^{\text{rel}}(t_A) + \frac{1}{c}T(t_A) + \frac{1}{c}I(t_A) . \quad (19.1)$$

The signal travel time is equal to the geometric range between the satellite and the receiver divided by the speed of light c . More precisely, the range is the difference between the antenna phase centers of the emitting and receiving antennas. The following equation for the signal travel time τ therefore contains the geometric range ρ_r^s , which refers to the satellite's center of mass and the receiving antenna's reference point and the additional term ξ_r^s contains the correction due to phase-center offsets of the transmitting and receiving antennas

$$\begin{aligned} \tau(t_A) &= \frac{\rho_r^s(t_A) + \xi_r^s(t_E, t_A)}{c} \\ &= \frac{\|\mathbf{r}^s(t_E) - \mathbf{r}_r(t_A)\| + \xi_r^s(t_E, t_A)}{c} , \end{aligned} \quad (19.2)$$

where $\mathbf{r}_r(t_A)$ is the receiving antenna's reference point at signal arrival time, and $\mathbf{r}^s(t_E)$ is the satellite's center of mass position at signal emission time. It should be noted that the term ξ_r^s also contains the contribution of code-phase patterns. These variable pseudorange delays are caused by the radiation pattern of the transmitting and receiving antennas, and thus depend on direction of signal transmission or reception as well as on frequency. Note that they are often referred to as group-delay variations in other literature [19.1, 2].

Since the local oscillator of the receiver is not synchronized with the GNSS time, the measured arrival time \tilde{t}_A as measured by the receiver deviates from the true arrival time due to the receiver clock offset $dt_r(t_A)$, instrumental delays d_r , and other errors like receiver noise and multipath $e_r^s(t_A)$

$$\tilde{t}_A = t_A + dt_r(t_A) + d_r + \frac{1}{c}e_r^s(t_A) . \quad (19.3)$$

The signal emitted by the satellite is affected by the clock offset of the onboard clock dt^s , the instrumental delays d^s , and the relativistic effect $\delta t_{\text{clk}}^{\text{rel}}$. Summarizing these terms leads to the following expression for the apparent emission time \tilde{t}_E

$$\tilde{t}_E = t_E + (dt^s(t_E) + \delta t_{\text{clk}}^{\text{rel}}(t_E)) + d^s . \quad (19.4)$$

The receiver measures the difference between the apparent arrival and emission time, which is converted with the speed of light to a pseudorange measurement

$$p_r^s(t_A) = c(\tilde{t}_A - \tilde{t}_E) . \quad (19.5)$$

Substituting (19.1)–(19.4) into (19.5) and introducing the signal identifier j to distinguish between different signals from the same satellite yields the expression for the pseudorange measurement

$$\begin{aligned} p_{r,j}^s(t) &= \rho_r^s(t) + \xi_{r,j}^s(t) + c(d_{r,j} - d_j^s) \\ &\quad + c(dt_r(t) - dt^s(t) + \delta t^{\text{rel}}(t)) \\ &\quad + I_{r,j}^s(t) + T_r^s(t) + e_{r,j}^s(t) . \end{aligned} \quad (19.6)$$

Equation (19.6) is the generic measurement equation for the pseudorange. The relativistic clock correction and the relativistic signal delay due to space-time curvature have been summarized in the single correction term $\delta t^{\text{rel}} = \delta t_{\text{stc}}^{\text{rel}} - \delta t_{\text{clk}}^{\text{rel}}$. The dependency on the frequency for the ionospheric delay and the dependency on the signal for the instrumental delays and multipath are indicated by the additional signal index j . Note that the indices for arrival and emission time have been dropped for brevity [19.3, p. 410], [19.4, p. 148].

However, it must be kept in mind that the difference between t_A and t_E is crucial to yield to correct geometric range. The observations are usually available with time tags referring to the measured arrival time. However, for the computation of the geometric range, the satellite position at true emission time t_E is required (19.2). In practice, the emission time can be approximated from $t_E = t_A - \tau$, where τ is computed from $\tau = \|\mathbf{r}^s(t_A) - \mathbf{r}_r(t_A)\|/c$ based on an initial guess for the receiver position. With an updated receiver position from a positioning solution, a more refined value for τ can be calculated in further iterations (Chap. 21).

Another important effect to remember in the modeling of the geometric term ρ_r^s is the rotation of the Earth-Centered Earth-Fixed reference frame during signal propagation. This effect must be accounted for if the receiver and satellite positions are computed in a noninertial reference frame. It is referred to as Earth rotation correction or Sagnac correction. One possibility to correct for this rotation is to rotate the satellites' positions backwards around the Earth rotation axis by an angle of $\tau \cdot \omega_{\oplus}$, where ω_{\oplus} is the rotation rate of the Earth [19.4]. Equivalently, the Sagnac correction can be applied as a correction

$$\Delta \rho_r^s = \frac{1}{c}(\mathbf{r}_r(t_A) - \mathbf{r}^s(t_E)) \cdot (\omega_{\oplus} \times \mathbf{r}_r(t_A)) \quad (19.7)$$

to the range between user and satellite, where ω_{\oplus} is the Earth's rotation vector, the operator \cdot denotes the inner product, and the operator \times denotes the vector product. Introducing the vector $\mathbf{S} = \frac{1}{2}(\mathbf{r}^s(t_E) \times \mathbf{r}_r(t_A))$ perpendicular to the user and satellite position vectors, the correction can also be expressed as [19.5]

$$\Delta \rho_r^s = \frac{2}{c}\mathbf{S} \cdot \omega_{\oplus} . \quad (19.8)$$

The different parameters in observation equation (19.6) may either be estimated, corrected for based on models, eliminated by combining observations (cf. Chap. 20), or even neglected depending on the application and the desired accuracy. For pseudorange-based positioning applications, the user may simply only estimate the receiver position \mathbf{r}_r , which is hidden inside the term ρ_r^s , and the receiver clock offset dt_r . Satellite clock offsets and positions, relativistic corrections, atmospheric delays and biases will instead be corrected for by external data or models. A GNSS service provider, on the other hand, will need to estimate the satellite's orbits, clock offsets, and biases based on a receiver network in order to be able to provide these values to users for positioning. Other applications may require precise estimates of the ionospheric or tropospheric delays, and these parameters may be estimated then based on GNSS data from a receiver network with known station positions. These examples for the use of GNSS data demonstrate that the decision of which parameters in (19.6) to model and which to estimate strongly depends on the desired application.

19.1.2 Carrier-Phase Measurements

A receiver does not only provide measurements of the pseudorange, but also of the signal's carrier phase from its phase lock loop (PLL). The receiver generates a replica of the carrier signal, aligns it with the incoming carrier from the satellite, and measures the fractional phase shift of both signals. When the range between user and satellite changes by more than one cycle, the receiver counts the full cycles and thus provides a continuous measurement. Due to the short wavelength of the carrier phase of approximately 19–25 cm, depending on the frequency, the carrier-phase measurement is much more precise than the pseudorange measurement. This advantage comes at the expense that this observable cannot provide an unambiguous measurement of the apparent satellite-receiver range like the pseudorange. Contrary to the pseudorange, where the navigation data modulated onto the signal is utilized to obtain an unambiguous measurement, the integer number of cycles between the satellite and the receiver at the start of carrier-phase tracking remains unknown. The observation equation for the carrier-phase measurement $\varphi_{r,j}^s$ in units of length is

$$\begin{aligned}\varphi_{r,j}^s(t) = & \rho_r^s(t) + \zeta_{r,j}^s(t) + c(\delta_{r,j} - \delta_j^s) \\ & + c(dt_r(t) - dt^s(t) + \delta t^{\text{rel}}(t)) \\ & - I_{r,j}^s(t) + T_r^s(t) \\ & + \lambda_j(\omega_r^s(t) + N_{r,j}^s) + \epsilon_{r,j}^s(t). \quad (19.9)\end{aligned}$$

The carrier-phase measurement is subject to clock offsets and instrumental delays in the receiver and satellite as well as atmospheric delays, receiver noise, and multipath similar to the pseudorange. The terms for geometric range, clock offsets, and tropospheric correction in this equation are identical to (19.6). Similar to the pseudorange, the carrier-phase observation is affected by phase-center offset and variations, which depend on the antenna phase-pattern and the frequency. To emphasize that the correction terms differ for both observables, the symbol $\zeta_{r,j}^s$ has been used to denote the corresponding correction in (19.9). In addition to that, the instrumental delays of receiver $\delta_{r,j}$ and satellite δ_j^s are different. The sign of the ionospheric correction is negative for the carrier-phase measurements, which will be explained in depth in the corresponding section.

Two additional terms have appeared in (19.9): the phase wind-up correction ω_r^s , which accounts for a change in the measured phase in case of rotations of the antennas. Finally, an unknown integer number of cycles $N_{r,j}^s$ is present in the measurement equation and converted to units of length using the wavelength λ_j of the corresponding frequency. The combined effect due to receiver carrier-phase tracking noise and multipath is summarized in the residual error term $\epsilon_{r,j}^s$ [19.4, p. 141].

19.1.3 Doppler Measurements

The observed frequency of a signal from a navigation satellite differs from its nominal frequency due to the Doppler shift caused by the relative motion of receiver and satellite. Additionally, the receiver's or satellite's clocks may be affected by a frequency offset or drift. In the receiver's PLL, the phase discriminator drives the numerically controlled oscillator (NCO) to match the frequency and phase of a local carrier-phase replica and the received signal. To compensate for the Doppler effect caused by relative motion of receiver and satellite or frequency deviations in the receiver or satellite clock, the NCO's frequency must be adjusted to keep both phases synchronized. This frequency adjustment in the PLL is reported by the receiver as the Doppler measurement [19.3, p. 411], [19.4, p. 467].

For the derivation of the Doppler observation equation, only the geometric Doppler effect is considered and atmospheric propagation delay, clock frequency deviations, and relativistic effects are neglected for now. Assuming that a receiver and a satellite are moving with velocities \mathbf{v}_r and \mathbf{v}^s respectively, the relation

$$f_r = f^s \left(\frac{1 + \left(\frac{\mathbf{e} \cdot \mathbf{v}_r}{c} \right)}{1 + \left(\frac{\mathbf{e} \cdot \mathbf{v}^s}{c} \right)} \right) \quad (19.10)$$

between the received frequency f_r and the transmitted frequency f^s can be found, where \mathbf{e} is the unit line-of-sight vector from the user to the GNSS satellite. Assuming that $v_r \ll c$ and $v^s \ll c$, (19.10) is expanded in terms of the quantity \mathbf{v}^s/c . Neglecting all terms involving \mathbf{v}_r and \mathbf{v}^s in higher order than two yields the following expression for the relation of received and transmitted frequency [19.6, p. 9]

$$f_r = f^s \left[1 + \left(\frac{\mathbf{v}^s}{c} - \mathbf{e} \right) \cdot \frac{(\mathbf{v}^s - \mathbf{v}_r)}{c} \right]. \quad (19.11)$$

For the following derivations, the clocks' frequency deviations are now introduced. The frequency deviation of the receiver clock is denoted df_r . The transmitted frequency is also subject to a deviation $\delta f_{\text{clk}}^{\text{rel}}$ from its nominal value caused by relativistic effects as well as a frequency deviation df^s due to imperfections of the frequency standard itself. As a result, the following

expression for the measured frequency \tilde{f}_r and the transmitted frequency \tilde{f}^s can be found

$$\begin{aligned} \tilde{f}_r &= f_r + df_r \\ \tilde{f}^s &= f^s + df^s - f^s \delta f_{\text{clk}}^{\text{rel}}. \end{aligned} \quad (19.12)$$

Introducing the expression for the observed Doppler shift $D_r^s = f_r - \tilde{f}^s$ and substituting (19.11) and (19.12) leads to

$$\begin{aligned} D_{r,j}^s &= \frac{1}{\lambda_j} \left(\frac{\mathbf{v}^s}{c} - \mathbf{e} \right) \cdot (\mathbf{v}^s - \mathbf{v}_r) \\ &\quad + (df_r - df^s) + \frac{c}{\lambda_j} \delta f_{\text{clk}}^{\text{rel}}. \end{aligned} \quad (19.13)$$

The term \mathbf{v}^s/c is a line-of-sight correction due to satellite motion. Note that the Doppler is positive if satellite and receiver approach each other, i.e., when the range rate $\dot{\rho} = \mathbf{e} \cdot (\mathbf{v}^s - \mathbf{v}_r)$ is negative.

19.2 Relativistic Effects

The relativistic effects in GNSS can be split into two different categories: delays affecting the signal path due to the Earth's gravitational potential, and deviations of the satellite clock frequency due to relativistic effects. Both of these are discussed in more detail in Sect. 5.4.

The Shapiro effect is a delay of the satellite signal due to the presence of the Earth's gravitational field, which causes a propagation delay due to space-time curvature. The corresponding delay correction $\delta t_{\text{stc}}^{\text{rel}}$ is computed from

$$\delta t_{\text{stc}}^{\text{rel}} = \frac{2\mu}{c^3} \ln \left(\frac{||\mathbf{r}^s|| + ||\mathbf{r}_r|| + \rho_r^s}{||\mathbf{r}^s|| + ||\mathbf{r}_r|| - \rho_r^s} \right), \quad (19.14)$$

where μ is the Earth's gravitational constant [19.7, 8]. Assuming a user on the surface of the Earth, the maximum Shapiro effect is approximately 60 ps or 2 cm for medium Earth orbit (MEO) satellites of GPS, GLONASS, Galileo, and BeiDou. For BeiDou satellites on inclined geosynchronous orbits (IGSO), it is on the order of 70 ps.

The onboard clock will be affected by a relativistic clock correction term $\delta t_{\text{clk}}^{\text{rel}}$ caused by the motion of the satellite as well as the change in the gravitational potential. These effects caused by special and general relativity lead to a frequency offset of the onboard clock with respect to a ground-based clock. To mitigate this effect, the satellite's clock frequencies are actually offset from their nominal values. However, noncircular

satellite orbits cause deviations from the mean frequency offset. The effect due to the orbit eccentricity can be computed from

$$\delta t_{\text{clk}}^{\text{rel}} = -\frac{2}{c^2} \sqrt{a\mu} e \sin E, \quad (19.15)$$

where a is the semimajor axis, e is the orbit eccentricity of the satellite orbit, and E the eccentric anomaly of the satellite, an angle related to the position of the satellite on its orbit [19.9]. The same correction can also be formulated using the satellite's position and velocity vector \mathbf{r}^s and \mathbf{v}^s [19.8, 10]

$$\delta t_{\text{clk}}^{\text{rel}} = -\frac{2}{c^2} (\mathbf{r}^s \cdot \mathbf{v}^s). \quad (19.16)$$

The time derivative of (19.16) is the frequency deviation due to the orbit eccentricity, which is required for the Doppler observation (19.13). It can be found from

$$\delta f_{\text{clk}}^{\text{rel}} = \frac{2\mu}{c^2} \left(\frac{1}{a} - \frac{1}{||\mathbf{r}^s||} \right). \quad (19.17)$$

It becomes immediately obvious from (19.17) that the correction is zero if the orbit radius is equal to the semimajor axis [19.11].

The typical order of magnitude for the relativistic clock offset correction is on the order of nanosec-

onds. Assuming the maximum tolerated eccentricity of 0.02 for GPS, the maximum value for $\delta r_{\text{clk}}^{\text{rel}}$ is approximately 45.0 ns, which corresponds to about 13.5 m. The maximum frequency deviation $\delta f_{\text{clk}}^{\text{rel}}$ in this case amounts to about 6.0 ps/s or approximately 2.0 mm/s. For GLONASS satellites, which have smaller orbit eccentricities, $\delta r_{\text{clk}}^{\text{rel}}$ and $\delta f_{\text{clk}}^{\text{rel}}$ are on the order of 8.0 ns and 1.2 ps/s. It becomes obvious that for GNSS satellites on near-circular orbits, the relativistic frequency offset deviation can be safely neglected in modeling Doppler observations. Even for satellites of the Quasi-Zenith Satellite System (QZSS) on their elliptical inclined geosynchronous orbits, this correction does not exceed 20 ps/s or 6 mm/s.

Even on a circular orbit the satellite clock would travel through a varying gravitational potential, since the Earth is not a perfect sphere with homogeneous mass distribution. The most prominent effect at the height of GNSS satellites is caused by the Earth's oblateness, described by the term J_{20} of the gravitation field expansion series. Often referred to as J_2 -correction, the periodical part of this correction can be found from

$$\delta r_{\text{clk}, J_2}^{\text{rel}} = -J_2 \frac{3}{2} \frac{r_{\oplus}^2}{c^2} \sqrt{\frac{\mu}{a^3}} \sin^2 i \sin 2u, \quad (19.18)$$

where R_{\oplus} is the equatorial radius of the Earth, $J_2 = 1.083 \cdot 10^{-3}$, the orbit inclination i , and the argument of latitude u , which is defined as $u = \omega + f$, the sum of argument of perigee ω and true anomaly f . It should be noted that the J_2 -correction as formulated in (19.18) is an additional correction term to the

conventional eccentricity correction of (19.16) [19.8, 10].

Obviously, the magnitude of the J_2 -correction is proportional to the inclination and inversely proportional to the semimajor axis of the satellite orbit. For MEO satellites, the correction varies between ± 62 ps for Galileo satellites, which have the largest semimajor axis, and ± 100 ps for GLONASS satellites, which have the lowest orbit altitude and the largest orbit inclination. For BeiDou IGSO satellites, the J_2 -correction decreases down to only ± 36 ps. Depending on the satellite system, this correction varies between approximately 1–3 cm and may therefore become relevant for the modeling of observations, especially for characterization of the high-precision rubidium atomic frequency standards of the GPS Block-IIIF satellites and the Galileo passive hydrogen maser.

Finally, it will be emphasized that all relativistic clock corrections discussed here so far have only considered the satellite clock. Of course, for a receiver moving through the Earth's gravitational potential, relativistic effects also apply to the receiver clock. For example, a space-borne receiver on an eccentric orbit would be subject to a similar relativistic clock correction as described by (19.16). The receiver relativistic clock corrections affect all observations in the same way. Therefore, considering them in the observation model can safely be neglected since they are entirely compensated in the receiver clock offset estimation. Modeling only becomes relevant, for example, when the true behavior of the receiver clock is to be characterized, a rather special application, which has not been regarded here.

19.3 Atmospheric Signal Delays

On its way to users on or close to the surface of the Earth, GNSS signals have to traverse the atmosphere. At a height of approximately 1000 km the signals encounter the ionosphere, a layer of charged particles, which extends down to an altitude of about 50 km. Underneath, the troposphere, a layer of electrically neutral gases, is predominant. Both layers change the speed as well as the direction of travel of the signal. This effect is referred to as refraction, and it is characterized using the refractive index n , which is the ratio of the speed of light and the speed in the medium v

$$n = \frac{c}{v}. \quad (19.19)$$

Since the atmosphere is not a uniform layer, the refractive index changes along the signal path. This change of

n causes the signal to be bent in accordance with Snell's law. However, the travel time on the curved path is shorter than for the straight-line path according to Fermat's principle of least time. The travel time of the signal can be found from the integral of the refractive index along the signal-path $n(l)$ from satellite to receiver

$$\tau = \frac{1}{c} \int_s^r n(l) dl. \quad (19.20)$$

The delay with respect to a signal path in a vacuum is then found from

$$\Delta \tau = \frac{1}{c} \int_s^r (n(l) - 1) dl. \quad (19.21)$$

If the refractivity index depends on the frequency, the associated media is called *dispersive*. In dispersive media, signals on different frequencies are subject to different delays. Furthermore, the carrier of the signal and its modulation travel at different velocities, which is referred to as code-carrier divergence. Therefore, phase refractive index n_p for the carrier and a group refractive index n_g for the code are defined as

$$\begin{aligned} n_p &= \frac{c}{v_p} \\ n_g &= \frac{c}{v_g}, \end{aligned} \quad (19.22)$$

where v_p and v_g are the phase and group velocities respectively. Both indices are related via

$$n_g = n_p + f \frac{dn_p}{df}. \quad (19.23)$$

These expressions will become relevant for the derivation of the signal delays for pseudorange and carrier phase in the dispersive ionosphere [19.4, 12].

In the following sections, the different GNSS signal delays caused by ionosphere and troposphere are briefly described. Simple models for the range delays are introduced. For a detailed discussion of tropospheric and ionospheric delays the reader may refer to Chap. 6 as well as Chaps. 38–39.

19.3.1 Ionosphere

The ionosphere is a layer of the atmosphere that contains electrically charged particles. The Sun's ultraviolet radiation causes gas molecules to break up into free electrons and ions. The delay of the signal depends on the number of free electrons along its path. The total electron content (TEC) is defined as the number of electrons in a tube with a cross section of 1 m^2

$$\text{TEC} = \int_s^r n_e(l) dl, \quad (19.24)$$

where $n_e(l)$ is the electron density along the signal path integrated from the satellite to the receiver. The electron density is not uniformly distributed along the signal path in the ionosphere. The ionization process affects the various layers of the ionosphere in different ways. As a result, the electron density changes along the signal path. The maximum density of charged particles can be found in altitudes between 250–400 km, which coincides with the so called F2 region of the ionosphere [19.13].

In addition to the spatial variability, there is also a temporal variation of the number of free electrons caused by the different amount of solar radiation during day and night. In the illuminated parts of the ionosphere, ionization takes place and the electron density reaches a peak value at 14:00h local time. With decreasing sunlight the recombination of electrons and ions prevails, and the number of free electrons is reduced again. In addition to this diurnal effect, the number of electrons undergoes a seasonal variation due to different amounts of solar radiation received in summer and winter as well as a long-term variation depending on the eleven-year solar cycle. However, there are also unpredictable short-term effects due to irregular changes in solar activity and traveling ionospheric disturbances. The latter can cause rapid changes in the electron density in a geometrically confined area. An active ionosphere can cause scintillation effects, which are rapid fluctuations in the carrier phase, as well as signal fading and may cause a threat to GNSS tracking especially in polar and equatorial regions [19.4].

As mentioned earlier, the ionosphere is a dispersive medium in which the signal delay depends on frequency and phase and code of the same signal travel at different velocities. The refractive index for the carrier phase n_p can be approximated to first order as

$$n_p \approx 1 - \frac{40.3n_e}{f^2}, \quad (19.25)$$

where n_e is the electron density and f is the frequency of the signal. The signal delay of the carrier-phase measurements can then be found from

$$\begin{aligned} \Delta\tau_p &= \frac{1}{c} \int (n_p(l) - 1) dl \\ &= -\frac{40.3 \cdot \text{TEC}}{cf^2}. \end{aligned} \quad (19.26)$$

The group refractive index n_g is found by substituting (19.25) into (19.23)

$$n_g \approx 1 + \frac{40.3n_e}{f^2}. \quad (19.27)$$

Similar to the carrier-phase delay, the pseudorange delay can be found by integrating the electron density along the signal path

$$\begin{aligned} \Delta\tau_g &= \frac{1}{c} \int (n_g(l) - 1) dl \\ &= \frac{40.3 \cdot \text{TEC}}{cf^2}. \end{aligned} \quad (19.28)$$

Multiplication with the speed of light converts the delays $\Delta\tau_p$ and $\Delta\tau_g$ to units of length

$$I = -I_p = \frac{40.3 \cdot \text{TEC}}{f^2} . \quad (19.29)$$

It becomes obvious that the pseudorange delay I has an opposite sign compared to the phase delay I_p , which has already been noted in (19.6) and (19.9). Furthermore, the pseudorange delay decreases with increasing frequency and depends, to first order, linearly on the total electron content along the signal path, also referred to as slant TEC (STEC). The TEC is measured in TEC units (TECU) which are defined as 10^{16} electrons/m². On the GPS L1 frequency, one TEC unit accounts for an approximately 16.2 cm delay.

A simple model for the ionospheric delay can be derived from the assumption that the ionosphere is a thin shell surrounding the Earth with evenly distributed electron density. In this case, the total electron content can be modeled as the TEC in the vertical direction (VTEC) scaled by an obliquity factor to account for the increase of ionospheric path length for lower elevation angles.

Figure 19.1 depicts the geometry between user, satellite, and the ionospheric pierce point (IP), where the line-of-sight vector penetrates the ionospheric shell at the mean ionospheric height h_I . The total ionospheric pseudorange delay in terms of VTEC and zenith angle z' at the IP can be found as

$$I = \frac{1}{\cos z'} \frac{40.3 \cdot \text{VTEC}}{f^2} . \quad (19.30)$$

The zenith angle at the IP is related to the zenith angle z at the user position through

$$\sin z' = \frac{R_\oplus}{R_\oplus + h_I} \sin z . \quad (19.31)$$

where R_\oplus is the radius of the Earth. This leads to the final expression

$$I = \frac{1}{\sqrt{1 - \left(\frac{R_\oplus}{R_\oplus + h_I} \sin z \right)^2}} \frac{40.3 \cdot \text{VTEC}}{f^2} . \quad (19.32)$$

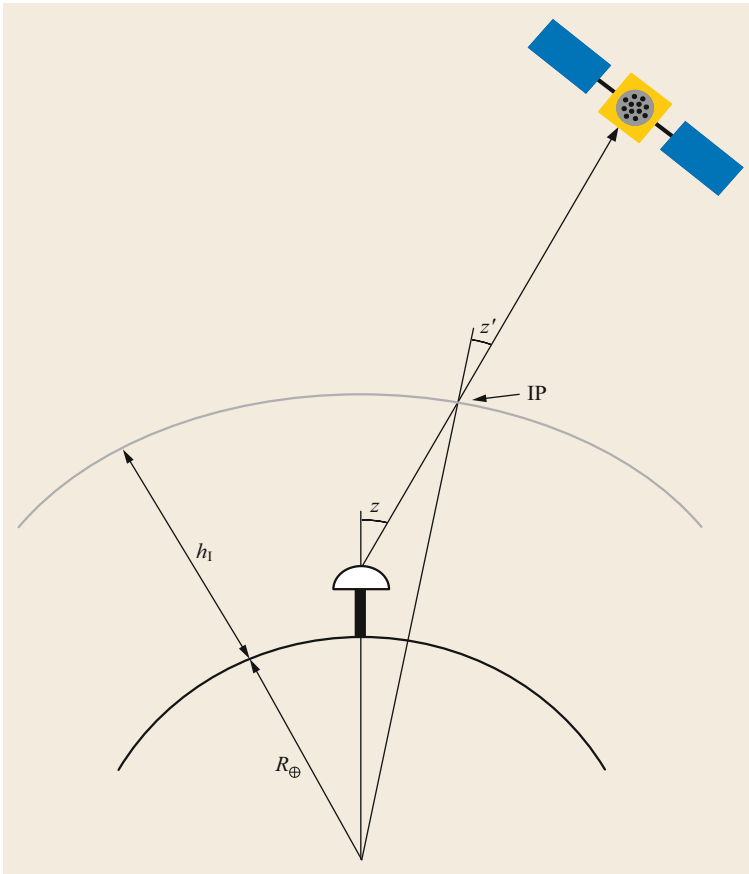


Fig. 19.1 Geometric relations between zenith angles z and z' at user position and ionospheric pierce point (IP) for the thin-shell ionospheric model (after [19.12])

The values for the obliquity factor vary between 1 and 3 for satellites at zenith and close to the horizon. Vertical TEC varies typically between a few TECU at local nighttime and several tens of TECU at local daytime. In the case of very high ionospheric activity, the peak value of VTEC can reach more than 200 TECU [19.4, 13].

The simple model introduced here can be utilized in different ways. With a correction value for the VTEC available, for example from a global or regional ionospheric map (GIM or RIM), users can apply (19.32) as a correction of the ionospheric delay in the observation model. In the case of single-frequency GPS, this method has been found to yield improved results compared to the broadcast ionospheric model [19.14, 15]. In the case of multifrequency receivers, the delay based on (19.32) can serve as an a priori correction, which allows a constrained estimation of a residual slant delay [19.16]. Attempts to model and estimate VTEC based on (19.32) have also been made [19.17].

19.3.2 Troposphere

The troposphere and the stratosphere together form the neutral atmosphere. This layer is nondispersive for signals in the L-band, which has a frequency range of 1–2 GHz. The L-band GNSS signals are therefore affected by the same delay in this part of the atmosphere irrespective of their center frequency, and the delay is also identical for pseudorange and carrier-phase measurements. The majority of the water vapor and moist gases are concentrated in the troposphere, the lower part of the neutral atmosphere. The stratosphere begins above the troposphere and extends up to a height of approximately 50 km. It contains mostly dry gases [19.12]. Both layers are usually referred to as troposphere in the GNSS nomenclature, and the same will be adopted for the remainder of this section.

The tropospheric path delay $\Delta\tau_T$ in units of time depends on the refractive index n integrated along the signal propagation path l

$$\Delta\tau_T = \frac{1}{c} \int_s^r (n-1) dl. \quad (19.33)$$

Note the similarity of this expression to (19.26) and (19.28). Using the tropospheric refractivity $N_T = 10^6(n-1)$ in (19.33) and converting the expression to units of length yields the total path delay

$$T_r^s = 10^{-6} \int_s^r N_T dl. \quad (19.34)$$

The total tropospheric refractivity N_T of air containing water vapor can be written as

$$N_T = k_1 \frac{p}{T} Z_d^{-1} + k_2 \frac{e}{T} Z_w^{-1} + k_3 \frac{e}{T^2} Z_w^{-1}, \quad (19.35)$$

where T is temperature, p and e are the partial pressures of the dry and wet constituents respectively, and Z_d and Z_w are the compressibility for dry and wet air respectively. The three constants are given as $k_1 = 77.6$, $k_2 = 64.8$ and $k_3 = 3.776 \cdot 10^5$ [19.18].

According to Saastamoinen [19.19] and Davis et al. [19.20], the total tropospheric delay can be separated into a hydrostatic (i.e., dry) part, which only depends on the pressure, and a nonhydrostatic part, which depends on the water vapor pressure profile and is also referred to as the wet part. The hydrostatic part is responsible for the majority of the delay. The nonhydrostatic part only accounts for a minor part of the total delay, however it is more difficult to model, since the water vapor content in the atmosphere can change rapidly.

If the atmospheric parameters, which determine the refractivity, were known for the entire signal path, the tropospheric delay could be obtained by integration of (19.34). In practice, however, p , T , and e may only be available with some effort from in situ measurements at the user position. A number of models have been developed, which relate the state of the atmosphere at an arbitrary height to the atmospheric parameters at the user height and thus allow the integration of (19.34) into zenith direction.

The resulting zenith tropospheric delay (ZTD) is then multiplied with a mapping function to yield the total delay along the signal path. The mapping function $m(E)$ may be different for the dry and wet delay, and depends on the satellite's elevation angle E as seen from the user. In its most general form, the delay can be written as

$$T_r^s = \text{ZTD}_d m_d(E) + \text{ZTD}_w m_w(E), \quad (19.36)$$

where ZTD_d and ZTD_w are the dry and the wet ZTDs, and m_d and m_w are the corresponding mapping functions.

One of the most fundamental models has been developed by Hopfield. It relates the hydrostatic and nonhydrostatic refractivities at the surface position to the refractivities at a given height above the user through a fourth-order function of height based on the assumption of a constant temperature lapse rate. The model also provides the corresponding mapping functions to compute the total slant delay. Atmospheric parameters are required as input [19.21].

Another fundamental model has been developed by *Saastamoinen* [19.19]. Based on his work, an expression for the hydrostatic delay has been derived that only depends on pressure, latitude and height above the geoid of the user position [19.20]. A consistent expression for the nonhydrostatic delay has been derived in [19.22], which depends on temperature, water vapor partial pressure, user location, and time.

Most of the mapping functions currently used to scale the zenith delay to a slant delay at satellite elevation angle E are based on a continued fraction expansion of $1/\sin(E)$, as defined by *Herring* [19.23] based on the suggestion of *Marini* [19.24]. This continued fraction expansion has the general form of

$$m(E) = \frac{1 + \left(\frac{a}{1 + \frac{b}{(1+c)}} \right)}{\sin(E) + \left(\frac{a}{\sin(E) + \frac{b}{(\sin(E)+c)}} \right)} \quad (19.37)$$

The mapping functions for wet and dry delay have identical form but differ in their values for coefficients a , b , and c . The coefficients for the wet mapping function according to *Herring* depend on latitude and height of the observation site as well as on local temperature. The dry mapping function only depends on latitude and temperature [19.23].

The Niell mapping functions (NMF) are also based on (19.37). The coefficients are obtained from a fit through radiosonde data from the Northern Hemisphere and are site- and time-dependent but do not require meteorological data [19.25]. Later, the accuracy of the NMF has been improved by deriving the coefficients of the mapping functions based on data of a numerical

weather model. The resulting isobaric mapping function (IMF) is more challenging to evaluate, since it depends on meteorological data at the epoch of the observations. This information is available from meteorological data centers for a global grid of reference points that are updated several times per day [19.26]. A similar approach has been done for the Vienna mapping function 1 (VMF1) [19.27].

To overcome the dependency on actual in situ atmospheric measurements, research has been conducted into developing globally valid models to provide the meteorological data for the tropospheric modeling. One of these models has been developed at the University of New Brunswick (UNB). The UNB3 model and the refined version UNB3m provide tabulated data to compute the temperature, pressure, and water vapor partial pressure based on latitude and time. These values are then used to compute the zenith delay based on the *Saastamoinen* model and mapped to a slant delay using the Niell mapping functions [19.28, 29]. In an alternative integrated model, the meteorological parameters provided by the global pressure and temperature model (GPT) [19.30] are used to compute wet and dry zenith delays, which are then mapped using the global mapping functions (GMF) [19.31] to the slant delay.

The UNB3m model yields a hydrostatic zenith delay of 2.3 m for a user on the meridian at the equator on day-of-year 120. The corresponding nonhydrostatic delay amounts to 0.27 m. Tropospheric mapping functions typically increase gradually from unity at zenith to values of about 2 at 30° and about 5 at 10° elevation angle. Below, the mapping function increases steeply to approximately 14–16 at 3°.

For high-precision applications, however, it may not be sufficient to only model the meteorological parameters with empirical data due to the temporal variation of water vapor in the atmosphere. It is common practice in geodetic processing to apply a priori corrections of the tropospheric delay in the modeling of the observations and include an additional zenith delay correction in the estimation parameters [19.12].

19.4 Carrier-Phase Wind-Up

This section describes the physical background of the wind-up effect of circular polarized waves, and presents the mathematical model for describing the effect as a function of the relative attitude of transmitter and receiver antennas. Since knowledge about the satellite's orientation is crucial for the computation of the wind-up effect, different satellite attitude models are presented.

19.4.1 Wind-Up Effect for Radio Waves

Navigation satellites typically emit right-hand circularly polarized (RHCP) waves (Chap. 4). The polarization of the wave determines how the electrical field vector changes, when the wave propagates from the emitter to the receiver. Circularly polarized waves consist of two signal components created by two crossed

dipoles \mathbf{x} and \mathbf{y} , which are perpendicular to each other, and create a sine wave with a relative phase shift of 90° . The resulting electrical field of the electromagnetic wave rotates in the x/y -plane.

Rotating the receiving antenna about its boresight direction, perpendicular to the two dipoles, changes the relative orientation to the electrical field, and thus the measured phase angle. Similarly, rotating the transmitting antenna changes the orientation of the instantaneous electrical field at the receiving antenna. This effect is referred to as phase wind-up or phase wrap-up. A full rotation of either the receiving or the transmitting antenna around its boresight vector causes a change in the measured carrier phase by one cycle for all frequencies. The corresponding range error is approximately 19 cm for L1 and about 25 cm for the L5 frequency [19.32].

Assuming that \mathbf{k} is a unit vector pointing from the transmitter to the receiver, an effective dipole vector \mathbf{D} can be defined in terms of the dipole unit vectors \mathbf{x} and \mathbf{y} as well as \mathbf{k} as

$$\mathbf{D} = \mathbf{x} - \mathbf{k}(\mathbf{k} \cdot \mathbf{x}) + \mathbf{k} \times \mathbf{y} . \quad (19.38)$$

A similar expression for the effective dipole vector \mathbf{D}' of the transmitting antenna can also be found

$$\mathbf{D}' = \mathbf{x}' - \mathbf{k}(\mathbf{k} \cdot \mathbf{x}') - \mathbf{k} \times \mathbf{y}' , \quad (19.39)$$

where \mathbf{x}' and \mathbf{y}' are the corresponding dipole unit vectors of the transmitting antenna. These properties are illustrated in Fig. 19.2. The phase wind-up ω in terms of cycles can then be found from

$$\omega = \text{sign}(\gamma) \arccos \left(\frac{\mathbf{D}' \cdot \mathbf{D}}{\|\mathbf{D}'\| \|\mathbf{D}\|} \right) , \quad (19.40)$$

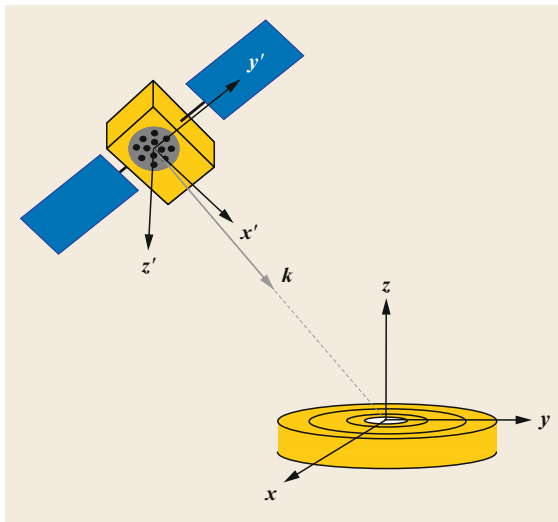


Fig. 19.2 Carrier-phase wind-up

with the term γ defined as

$$\gamma = \mathbf{k} \cdot (\mathbf{D}' \times \mathbf{D}) . \quad (19.41)$$

Note that continuity of the phase correction in (19.40) has to be maintained if the relative rotation exceeds 360° .

The two different contributions due to satellite or receiver antenna rotation manifest themselves differently in the observations. The phase wind-up effect due to rotation of the receiving antenna is identical for all received signals. If uncorrected, it will therefore be compensated by an epoch-wise estimation parameter, which is common for all carrier-phase observations, for example a receiver clock estimate. On the other hand, this effect may also be exploited for spin-rate estimation of the receiving antenna [19.33]. The phase wind-up contribution from the transmitting antenna is different for each satellite, however, and is typically included in the modeling of undifferenced observations. For differential processing, the residual contribution is on the order of a few millimeters for short baselines of several hundreds of kilometers or less. However, residual errors of a quarter of a wavelength can occur in double differences of observations for baseline lengths of thousands of kilometers [19.34].

The previous discussion has assumed the ideal case, in which only right-hand circularly polarized (RHCP) signals have been considered. In reality, a fraction of the transmitted signal is also left-hand circularly polarized. However, this effect can generally be neglected for users on or close to the Earth, including satellites on low-Earth orbits within a cone of 16° of the transmitting boresight [19.34].

19.4.2 GNSS Satellite Attitude Modeling

Knowledge of the navigation satellite's attitude is obviously crucial for the computation of the phase wind-up effect due to rotation of the transmit antenna. The orientation of a navigation satellite is determined by two requirements. Firstly, the boresight vector of the transmit antenna must be pointed towards the Earth. Secondly, the solar panels have to be oriented to the Sun in order to provide sufficient power for the spacecraft. To fulfill the first requirement, the satellite's body-fixed z -axis, which points parallel to the antenna boresight vector, is always oriented towards the center of the Earth. The second requirement is achieved by rotating the satellite around its z -axis in order to keep the solar panel axis perpendicular to the Sun's direction. This continuous change in orientation during the orbital period is referred to as yaw steering (Sect. 3.4). The

rotation matrix \mathbf{R}_{YS} from the satellite body-fixed coordinate frame to the Earth-Centered Earth-Fixed (ECEF) is then [19.35]

$$\begin{aligned} \mathbf{e}_z &= -\frac{\mathbf{r}}{\|\mathbf{r}\|} \\ \mathbf{e}_y &= \mathbf{e}_z \times \mathbf{e}_\odot \\ \mathbf{e}_x &= \mathbf{e}_z \times \mathbf{e}_y \\ \mathbf{R}_{YS} &= [\mathbf{e}_x \ \mathbf{e}_y \ \mathbf{e}_z]. \end{aligned} \quad (19.42)$$

Here, \mathbf{r} is the satellite position vector, \mathbf{e}_\odot is the unit vector pointing from the satellite to the Sun, and \mathbf{e}_x , \mathbf{e}_y , and \mathbf{e}_z are the body-fixed unit vectors as depicted in Fig. 19.3.

The yaw angle Ψ is the angle between the body-fixed x -vector and the along-track vector pointing towards the general direction of the satellite's velocity vector. The yaw angle Ψ and the yaw rate $\dot{\Psi}$ can be found from

$$\begin{aligned} \Psi_{YS} &= \arctan\left(\frac{-\tan \beta}{\sin \mu}\right) \\ \dot{\Psi}_{YS} &= \dot{\mu} \frac{\tan \beta \cos \mu}{\sin^2 \mu + \tan^2 \beta}, \end{aligned} \quad (19.43)$$

where μ is the orbit angle measured from orbit midnight and β is the elevation of the Sun above the orbital plane [19.35].

However, this attitude law is subject to a singularity that occurs if the Sun incident angle β is zero.

The yaw angle is therefore undefined twice along the spacecraft orbit in this case: when the satellite is closest to the Sun at orbit noon or furthest away from the Sun at orbit midnight. Furthermore, this attitude law requires the satellite to perform an instantaneous rotation by 180° after crossing the orbit midnight or noon position [19.35]. In reality, the maximum possible yaw rate of the spacecraft is already exceeded for small β -angles due to hardware limitations of the satellite's attitude control system and will cause the actual yaw angle to differ from the attitude law in (19.43). The resulting yaw attitude profiles differ between miscellaneous constellations, and even between different satellite types within the same constellation.

The Block II/IIA satellites of the GPS constellation cannot follow the nominal yaw rate defined by (19.43) close to orbit noon if $|\beta|$ is less than 3.6 – 4.9° . The satellite rotates with the maximum yaw rate of 0.10 – $0.13^\circ/\text{s}$ during these phases and the actual yaw attitude lags the required yaw angle. When the satellite proceeds on its orbit after crossing the orbit's noon position, the required yaw rate decreases and the satellite can catch up again and reaches nominal attitude. If a Block II/IIA spacecraft enters the Earth's shadow close to orbit midnight, its attitude control system cannot determine the yaw orientation anymore, since it solely relies on the Sun sensors for this task. The satellite starts a yaw motion with the maximal rate in this case until it leaves the eclipse again. It then recovers its nominal attitude. The direction of the yaw motion in the eclipse phase is

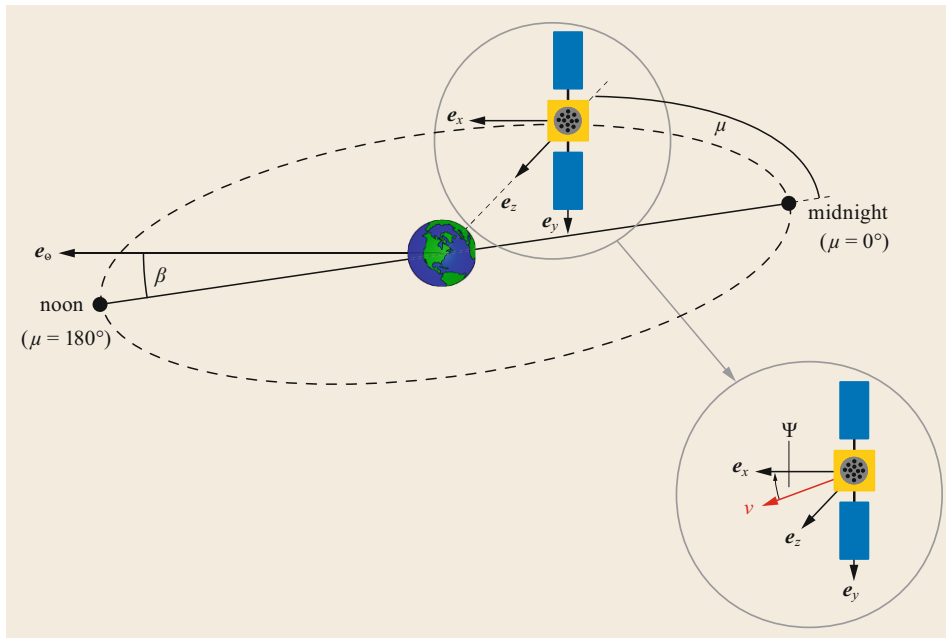


Fig. 19.3 Orbit of a GNSS satellite depicting the orbit angle μ with the orbit noon and midnight positions respectively, the Sun incidence angle β , and the yaw angle Ψ

predictable since a modification of the attitude control system was implemented in the mid 1990s. However, the yaw turn direction during the post-shadow recovery maneuver cannot be predicted reliably [19.35]. The newer generation of Block IIR satellites are capable of maintaining the nominal attitude even during eclipse phase but are still limited by a maximum yaw rate of $0.20^\circ/\text{s}$, which is reached for $|\beta| < 2.4^\circ$ [19.36]. The midnight and noon turn maneuvers of the newest Block-IIF satellites differ again from the previous satellites. As soon as the satellite enters the Earth's shadow, it starts a rotation with a constant rate, which is selected such that it brings the satellite to the required attitude at shadow exit. Errors of a few degrees with respect to the nominal orientation are corrected with a short recovery maneuver. For noon turn maneuvers with $|\beta| < 4^\circ$, the satellite's rotation is limited to a constant yaw rate of $0.11^\circ/\text{s}$. The difference between nominal and actual yaw attitude may amount to $\pm 180^\circ$ or $\pm 90^\circ$ for noon and midnight turn maneuvers respectively [19.37, 38].

GLONASS satellites are likewise affected by the need for special orbit noon and midnight maneuvers. However, the implementation of these maneuvers differs significantly from GPS. The current GLONASS-M satellites perform a midnight turn maneuver directly after shadow entry. The satellite starts a rotation with the maximum hardware yaw rate of approximately $0.24\text{--}0.26^\circ/\text{s}$ until it has reached the nominal orientation required at shadow exit. It then maintains this constant orientation throughout the eclipse phase and follows the yaw-steering attitude law after leaving the Earth's shadow. As a result, the actual yaw angle may precede the nominal angle by up to $\pm 180^\circ$. For the noon turn maneuvers, the nominal yaw rate exceeds the maximal hardware yaw rate for $|\beta|$ less than 2° . Again, the satellite rotates with its maximum yaw rate during this maneuver. In contrast to GPS satellites though, the rotation already starts before the nominal yaw rate exceeds the maximal possible rate. As a result, the actual

yaw orientation first precedes and then lags the nominal one [19.39].

The Galileo In-Orbit Validation (IOV) satellites also follow the yaw steering law for $|\beta| < 2^\circ$. In this case, the attitude control system uses a patented *dynamic yaw steering* concept. The idea behind this concept is to use a modified Sun angle as input for the control law in (19.43), in order to ensure that the maximum hardware yaw rate is not exceeded [19.40, 41].

The first QZSS satellite Michibiki on its elliptical inclined geosynchronous orbit (IGSO) uses a completely different attitude control approach for small β -angles. Already at $|\beta| < 20^\circ$, the satellite switches from yaw-steering attitude mode to orbit-normal mode, in which the satellite's body-fixed axes are aligned with the orbital frame. As a result, the yaw angle is 0 or 180° , depending on the convention, and the yaw rate is zero [19.42]. This switch between the two modes is performed as a rotation with a constant yaw rate. When the β -angle approaches the threshold, the switch typically takes place when the actual attitude is closest to the orientation in the other mode, which happens once per orbit (i. e., once per day). The attitude differences for small β -angles between the yaw-steering and the orbit-normal mode can be up to $\pm 180^\circ$ [19.43].

The BeiDou constellation consists of satellites on geostationary Earth orbits (GEOs), medium-Earth orbits (MEOs), and IGSOs. The GEO satellites use orbit-normal attitude exclusively [19.44]. The MEO and IGSO satellites are in yaw-steering mode except for small β -angles [19.45]. Based on the time intervals for orbit normal mode of two IGSO satellites provided in [19.46], the threshold value for the attitude mode switch seems to be close to $\pm 4^\circ$.

Satellite attitude modeling differs quite significantly among the various constellations and also among the satellite types in the same constellation. Exact modeling of the attitude is crucial though for many high-precision applications, as an attitude error of 90° already amounts to a carrier-phase error of 0.25 cycles.

19.5 Antenna Phase-Center Offset and Variations

In the previous sections, the range to be measured has been referred to as the distance between satellite and receiver. Using a more precise formulation, it is, of course, the distance between the electrical phase centers of the transmitting and receiving antennas. For high-precision applications, however, the introduction of the electrical phase center as a single point that all signals refer to is no longer possible, since variations depending on azimuth, elevation, and frequency occur. This effect

exists for code and phase observations. The following sections will introduce the concepts for modeling these effects and briefly explain calibration methods used to compute corrections for it. Finally, examples illustrating the magnitude of the correction will be presented, and the Antenna Exchange (ANTEX) format commonly used to disseminate the corrections will be introduced. For additional information an antennas the reader may also be referred to Chap. 17.

19.5.1 Overview

As already mentioned in the introduction, modeling the origin of all signals transmitted or received by an antenna as a single point may not be sufficient for high-accuracy applications due to variations depending on the direction of the incoming signals as well as their frequency. However, through calibration these effects can be removed in the modeling of the observations, which is the reason for the appearance of the correction terms $\xi_{r,j}^s$ in (19.6) and $\zeta_{r,j}^s$ in (19.9). The following introduction will mainly focus on the calibration of carrier-phase-center offsets and variations. However, similar methods have also been applied recently to calibrate the code-phase patterns that affect the pseudorange observations [19.47].

Two applications have mainly motivated the efforts for antenna calibrations [19.48]. The first is real-time kinematic (RTK) positioning with different antenna types. Using identical receiver antennas, the modeling errors due to phase-center offsets and variations cancel out completely in differential processing with short baselines. With different antennas this error cancellation is no longer possible and may adversely affect ambiguity resolution [19.49–51]. The second application is the processing of global network solutions,

where omitting phase-center offsets and variations of transmitting and receiving antennas in the modeling has an effect on estimates of station positions and tropospheric delay [19.52–54].

Figure 19.4 shows a schematic of a GNSS choke-ring antenna with radome to illustrate the modeling of phase-center offsets and variations. The antenna reference point (ARP) is introduced as a common physical point for all signals. It is defined by the International GNSS Service (IGS) for most antennas as the intersection of the vertical symmetry axis of the antenna with the bottom of the ground plane or choke ring. The mean electrical phase center differs from the ARP by the phase-center offset vector $\mathbf{r}_{\text{PCO},j}$. Note that this offset is frequency dependent. The phase-center offset correction is computed as its projection on the negative line-of-sight vector for the receiving antenna

$$\zeta_{\text{PCO},r,j} = -\mathbf{e} \cdot (\mathbf{A} \mathbf{r}_{\text{PCO},r,j}) . \quad (19.44)$$

The matrix \mathbf{A} is a direction cosine matrix, which transforms the phase center offset (PCO) vector, given in an antenna-fixed coordinate system, to the coordinate system, in which \mathbf{e} is provided, typically the Earth-Centered Earth-Fixed (ECEF) frame. In the case of a static antenna, for example at a geodetic reference

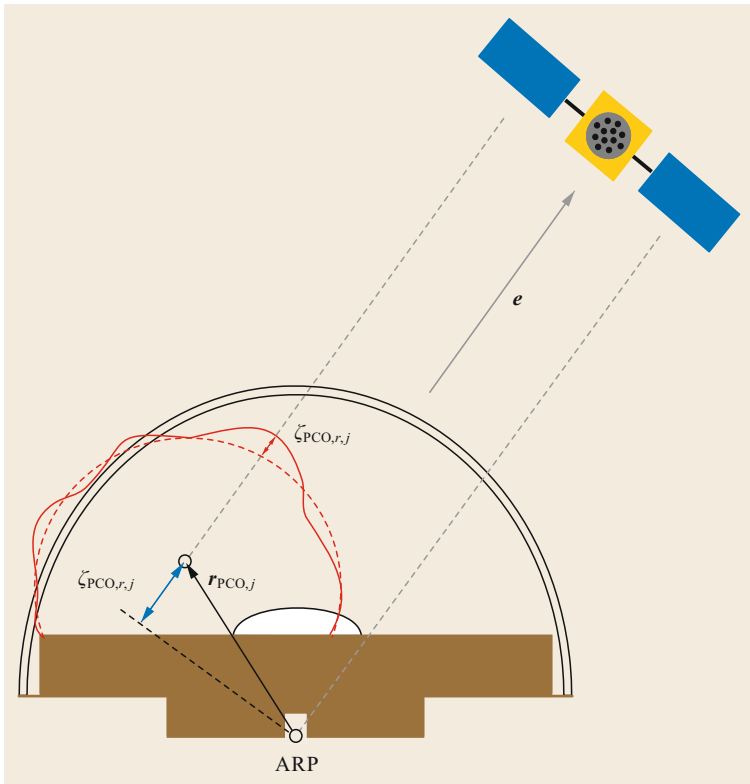


Fig. 19.4 Schematic of corrections for phase-center offset (blue) and phase-center variation (red) of a receiving choke-ring antenna with radome. The solid red line indicates the elevation- and azimuth-dependent phase-center variations with respect to a reference wavefront (dotted red line) (after [19.12])

station, the matrix \mathbf{A} is constant. In other cases, like airborne or space-borne applications, the attitude of the airplane or satellite may change continuously and must be taken into account for proper modeling of the phase-center offset correction.

For the transmitting antenna, the satellite's phase-center offset $r_{\text{PCO},j}^s$ has to be applied. Since transmitting and receiving antenna phase-center offsets are defined using the same convention, the satellite PCO must be projected on the positive line-of-sight vector. The matrix \mathbf{A} now describes the rotation from the satellite antenna coordinate system to the coordinate system choice of e

$$\zeta_{\text{PCO},j}^s = e \cdot (\mathbf{A} r_{\text{PCO},j}^s). \quad (19.45)$$

Only including the mean antenna phase center in the modeling would still leave a residual error for an individual measurement. Therefore, scalar corrections for the phase-center variation of the transmitting antenna $\zeta_{\text{PCV},j}^s$ and the receiving antenna $\zeta_{\text{PCV},r,j}$ are applied, which may be frequency-, elevation-, and azimuth-dependent [19.48]. The complete correction is then found from

$$\zeta_{r,j}^s = \zeta_{\text{PCO},j}^s + \zeta_{\text{PCO},r,j}^s + \zeta_{\text{PCV},j}^s + \zeta_{\text{PCV},r,j}. \quad (19.46)$$

19.5.2 Calibration Techniques

The two main techniques used to find the calibration for the phase-center offset and variations are discussed in this section. The first technique is referred to as relative antenna calibration. In this case, a reference and a test antenna are mounted on concrete pillars at a distance of 5 m. The reference antenna must always be of identical type and its phase-center offset and variations are assumed to be zero. In the case of the calibration procedure carried out by the US National Geodetic Survey (NGS) for IGS, an Allen Osborne Associates Dorne Margolin T antenna is selected as a reference. Using differential carrier-phase observations of live signals, the mean relative phase-center offsets are determined for each frequency. The phase-center variations are then calibrated with respect to the previously estimated phase-center offsets, again independently for each frequency. In the NGS calibrations, only a dependency on elevation is considered [19.51]. Note that the satellite's phase pattern does not affect this calibration procedure, as it cancels out in the single-difference observations.

The relative calibrations are suitable for processing RTK solutions in small networks, since the antennas involved observe a satellite at a similar elevation angle. For global solutions of large-scale networks, this assumption is no longer true. Therefore, calibrations

without the dependency on a reference antenna are necessary in this case. They are also referred to as absolute calibrations. Two different methods are available to obtain calibrations for receiving antennas: anechoic chamber tests and robot calibration with live signals.

During anechoic chamber tests, GNSS signals are generated and broadcast through a transmitting antenna to the antenna under test in an environment shielded against external radio-frequency signals and multipath. The test antenna is mounted in such a way that it can be rotated around two axes, to cover the desired range of azimuth and elevation angles. Measurements of phase delays are then obtained by comparing the transmitted and received carrier phase in a network analyzer. These phase delay measurements are afterwards used to estimate the antenna's phase-center offset and azimuth- and elevation-dependent pattern [19.48, 55].

For the second absolute calibration technique, the antenna under test is fixed to a mount that can rotate and tilt the antenna, which is located in an environment where live GNSS signals can be observed. Of course, the test environment cannot be considered to be free of multipath, which will affect the phase-center calibration. For the elimination of these errors, the repetition of the satellite geometry can be exploited. Taking measurements of two different days with repeated satellite geometry in an identical antenna environment will yield identical multipath errors in both measurements. Taking the time difference of these measurements eliminates not only the multipath but unfortunately also the phase-center variations to be calibrated. The solution to this problem is to rotate the test antenna on the second day. The rotation leads to a different phase-center variation effect in the observation but the multipath error and the satellite's antenna phase-center correction are still the same and can be eliminated through the time difference. The differences in tropospheric and ionospheric delays between the two days are eliminated by using measurements of an additional antenna at a short baseline. Because the phase pattern is derived from differential measurements, a common offset in all phase center variations (PCVs) is also eliminated. Since it is identical for all measurements on the same frequency, it can be compensated for in the processing through the clock offset or a hardware delay. Still, the calibration method is referred to as *absolute* calibration, since it is independent of the phase pattern of a reference antenna [19.56, 57].

Further refinements to the previous method have now led to the use of a robot which rapidly rotates and tilts the antenna. In this way the multipath can be removed without the need for a repeating constellation geometry. Through an automated procedure, several thousands of different antenna orientations can be achieved. As a result, the entire hemi-

spherical field of view of the antenna can be evenly covered with measurements and the phase pattern can be determined with high resolution. Furthermore, elevation- and azimuth-dependent corrections can be derived [19.58]. A disadvantage of this technique is that signals on new frequencies cannot easily be calibrated if they are not yet emitted by a sufficient number of satellites.

PCOs and PCVs not only affect GNSS observations at the receiving antenna, but also at the satellite's transmitting antenna. Therefore the observations must also be corrected for the satellite-dependent part of the phase pattern for processing of long baselines or undifferenced observations. Even though attempts have been made to perform absolute robot calibrations of a spare Block II/IIA satellite antenna [19.59], mean offsets and phase patterns are estimated from GNSS data of a global receiver network. The effect of the ionosphere must be removed for the estimation of the PCO and PCV corrections with the ionosphere-free combination of observations. Therefore, the derived parameters do not refer to the phase centers and variations of the individual signals. In the current processing, azimuth dependency in the satellite phase patterns are neglected

and only block-specific patterns are estimated. However, significant differences in the mean phase-center offsets in the antenna boresight direction have been identified even between satellites of the same type. Therefore, PCOs are estimated independently for each transmitting antenna [19.60].

Further information on the calibration of receiver and satellite antennas is provided in Sect. 17.6.2 of this Handbook.

19.5.3 Examples for Phase-Center Variations

The antenna exchange format (ANTEX) is a data format established within IGS for the dissemination of PCOs and PCV patterns. It allows the storage of data for transmitting and receiving antennas in ASCII format. The PCOs and PCVs can be stored independently for each frequency. The phase patterns can be stored as tabular values with selectable grid size for elevation and azimuth, or as rotational symmetric patterns without azimuth dependency [19.61].

As an example for a geodetic antenna, Fig. 19.5 depicts the PCV pattern of the Leica AR25 geodetic

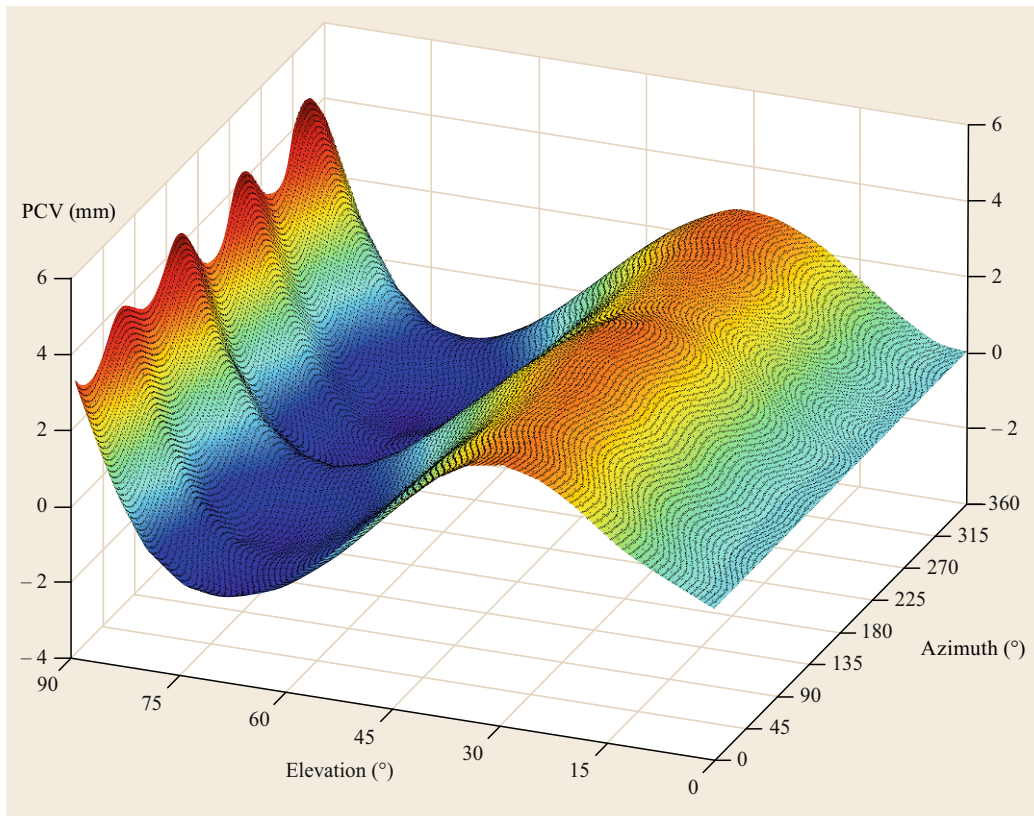


Fig. 19.5 Phase-center variation pattern for the GPS L1 frequency for a Leica AR25 antenna without radome

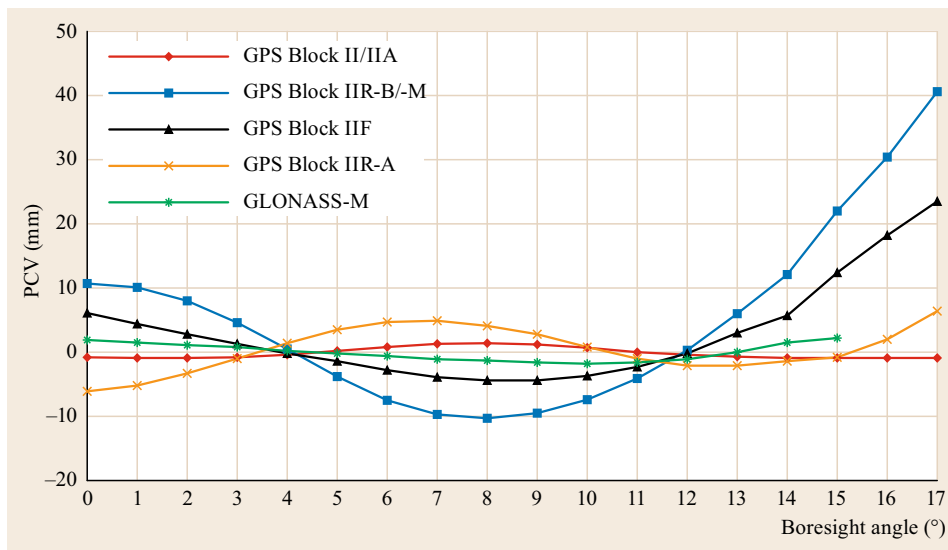


Fig. 19.6 Ionosphere-free phase-center variation patterns of GPS and GLONASS based on IGS data. GPS Block II and Block IIA satellites as well as Block IIR-B and Block IIR-M satellites have identical phase patterns. GLONASS-M PCVs are only available up to 15° boresight angle

antenna without radome. It becomes obvious that the pattern is normalized to zero at 90° elevation, since the absolute value of the PCV cannot be estimated. An azimuth-dependent pattern is also clearly visible for low elevations. Close to the antenna's horizon, the phase pattern exhibits peak-to-peak variations of approximately 2 mm

The plot in Fig. 19.6 depicts the ionosphere-free phase patterns for selected GPS and GLONASS satellites based on the block-specific calibrations by IGS. Since azimuth dependency of the phase pattern is not

considered in the IGS calibrations, the PCV variation is plotted over the boresight angle. The phase patterns for GLONASS satellites are only available for boresight angles up to 15°. For GPS satellites the phase patterns have been extended to 17° based on the observations from low Earth orbit (LEO) satellites. It becomes obvious that the phase patterns are identical for GPS Block II and Block IIA satellites as well as for Block IIR-B and Block IIR-M satellites. The maximum peak-to-peak phase pattern variation is found for the Block IIR-B/IIR-M satellites.

19.6 Signal Biases

This section discusses signal biases for pseudorange and carrier-phase observations. Most biases are not directly observable, which hampers the derivation of universally valid definitions. However, their origin will be explained in this section and examples for commonly used bias definitions will be provided.

19.6.1 Pseudorange Biases

The model of the pseudorange observation in (19.6) includes biases to account for the instrumental delays. Imperfect synchronization of the different signals is the cause of these delays and they occur at different stages along the signal path from generation to reception.

It is intuitive to separate the total bias into a receiver-dependent and a satellite-dependent term $d_{r,j}$ and d_j^s respectively, as in (19.6). The underlying assumption is that the receiver-dependent part is identical

for all signals of this type tracked by the same receiver. The satellite-dependent part is assumed to be equal for every receiver tracking the corresponding signal. Since the processing chains of the various signals differ, their instrumental delays may differ as well and the terms in (19.6) bear the signal index j to denote frequency and signal dependency. The biases are assumed to be time invariant, therefore an explicit dependency on time has been omitted in the notation. These assumptions are made here to simplify the discussion, but they are subject to restrictions that will be discussed at the end of this section.

The satellite-dependent part of the bias is caused by delay differences in the analog and digital paths of the signal generation unit and of the antenna, which are not identical for the various signals. It is important to note that the absolute delay of a single signal is unobservable. Only differential code biases (DCBs) become

observable by selecting one signal (or a combination of signals) and referencing the other signals to it. A differential code bias d_{BA}^s is defined as the difference between the absolute biases of the signals A and B

$$d_{BA}^s = d_A^s - d_B^s. \quad (19.47)$$

For each navigation system, the system time refers to a specific signal or to a signal combination, and the satellite clock correction terms refer to the deviation of the satellite clock from the system time. The absolute biases of the reference signal or combination are unobservable and lumped into the satellite clock correction terms. In the case of GPS, the ionosphere-free combination of the legacy P(Y) signals on L1 and L2 is used as a reference to compute the clock solution. If a user processes single frequency observations or a different combination of observations, the corresponding biases must be corrected for. For this purpose, each GPS satellite transmits a correction parameter, the so-called time group delay T_{GD} , which allows the correction of the bias between the ionosphere-free linear combination of L1/L2-P(Y) signals and the single-frequency P(Y) signals. The T_{GD} is expressed in terms of the L1/L2 P(Y) DCB as

$$T_{GD} = -\frac{f_{L2}^2}{f_{L1}^2 - f_{L2}^2} (d_{GC1W}^s - d_{GC2W}^s). \quad (19.48)$$

In this notation, the frequency index j has been replaced by a four-letter signal label, which consists of the three-letter receiver independent exchange (RINEX) v3 observation code preceded by a satellite system indicator, in this case G for GPS. With the help of the time group delay correction, the P(Y) single frequency observations can be aligned with the ionosphere-free clock reference. The newer generation of GPS satellites will also transmit so-called inter-signal corrections (ISCs), which contain the necessary DCBs for single-frequency positioning with C/A, L2C, and L5 code. In addition to the T_{GD} or ISC values provided by the navigation system itself, precise estimates of these values are also available from IGS analysis centers [19.62, 63].

The previous discussion has shown that significant effort is already necessary to process mixed signals, or combinations thereof, from a single navigation system. The situation becomes even more complicated if observations of different navigation systems are being used. In addition to the previously mentioned DCB corrections, which users may have to apply depending on the signal selection for processing, two different navigation systems are likely to exhibit offsets between their system timescales as well. It lumps together into an overall inter-system bias (ISB) [19.64–66] that should be estimated as an additional parameter by the user when

processing more than one constellation at a time. Alternatively, it may partly be mitigated by correction values for system-time offsets provided by modernized navigation systems as part of the broadcast data. Consistent estimation, dissemination and application of this variety of corrections remains as a future challenge for GNSS system operators, external service providers and users.

Different signal paths from the antenna to the correlator in the receiver are responsible for receiver-dependent instrumental delays. These delays often have received lesser attention, since common biases for all observations are simply lumped into the receiver clock estimate. However, in the case of multi-GNSS processing, the receiver-dependent instrumental delays of signals from different GNSSs are not likely to be identical. At this point it becomes important again to estimate this offset as one parameter per additional GNSS or use external corrections.

The pseudorange biases of the legacy signals of the GLONASS system pose an additional challenge. Contrary to most other modern navigation signals, the legacy GLONASS signals on the L1 and L2 band rely on the frequency division multiple access (FDMA) technique. These signals are not transmitted on the same frequency as the code-division multiple access signals (CDMA) but instead on a range of slightly shifted subbands close to the L1 and L2 center frequencies. The receiver distinguishes the signals by tracking on different subbands. In the case of GLONASS, 14 such subbands exist and are shared by two satellites placed on antipodal orbital positions.

It has been recognized already very early in the utilization of GLONASS in combination with GPS that the GLONASS pseudorange and carrier-phase observations are affected by different inter-channel receiver hardware biases. They are caused by filters as well as components of the signal processing chain such as antennas, low-noise amplifiers (LNAs) and cables, which exhibit different delays depending on frequency. These biases may vary even between identical receiver types due to the variability of component characteristics and temperature dependencies [19.67] and thus hamper the separation into a satellite- and receiver-dependent parts.

The remainder of this section shall be dedicated to discuss the limitations of the previous assumptions on the pseudorange biases, namely their invariability with time and their separability into purely receiver- and satellite-dependent parts. The constantness of a signal bias will in practice be limited, for example, by the fact that electrical components change their properties with temperature. Temperature variations inside the receiver [19.68, 69] but also inside the GNSS satellite [19.70] may cause variations of

signal biases. If biases can be assumed constant and over which timescales this assumption is valid very much depends on the application as well as local conditions at the receiving and transmitting equipment.

The second assumption that needs restrictions is the separability of the instrumental delays into the receiver- and satellite-dependent parts also for the CDMA signals. For the example of GPS, close inspection of the interface control document (ICD; [19.71]) reveals that the T_{GD} and ISCs provided through the broadcast data are only valid for a specific receiver type with clearly defined bandwidth as well as correlator type and spacing. The underlying reason is that different correlator implementations may respond differently to possible satellite-specific distortions of the transmitted waveforms. As a result of the signal distortion, the instrumental delays inside the receiver will vary but the change may be different for each satellite and thus cannot be compensated for by a common change in the receiver-dependent signal delay. Care must therefore be taken while estimating and applying satellite-dependent bias corrections to receivers with different correlator settings. This especially holds true for high-performance multipath-mitigation techniques, which are often based on very narrow correlator spacings [19.72, 73].

19.6.2 Carrier-Phase Biases

Similar to the pseudorange, the carrier-phase observations are also affected by instrumental delays. These biases are again split up into a satellite- and receiver-dependent part, and are denoted δ_j^s and $\delta_{r,j}$ in (19.9). Carrier-phase biases are a particular nuisance, since

their fractional part prohibits the integer resolution of the unknown ambiguity N . In precise positioning (Chaps. 25 and 26), these biases can be eliminated directly by processing double differences of observations using a reference station. Alternatively, the carrier-phase biases can be eliminated with correction values computed from a global or regional reference station network. A universally valid definition of the correction values cannot be provided, since it strongly depends on the parametrization of the estimation system. However, significant effort has been made to characterize the satellite-, and receiver-dependent parts of the carrier-phase biases and enable the provision of carrier-phase bias corrections to allow integer ambiguity resolution [19.74–79].

Similarly to the legacy GLONASS pseudorange observations, also the biases of the carrier-phase measurements of these signals are subject to further complications. As already discussed in the previous section, the employed FDMA signal scheme requires the GLONASS signals to be tracked on different frequencies. This causes decimeter-level inter-frequency biases in the carrier-phase measurements. These biases have been found to have a linear dependency on the frequency and they are approximately equal for L1 and L2, if expressed in units of length. However, the biases have been shown to exhibit significantly different slopes depending on receiver type. Therefore cancellation is not possible in differential processing with different receiver brands [19.80]. It has been demonstrated that the main part of the bias is not created in the analog part of the receiver but in the digital signal processing (DSP) chain. These biases are therefore not dependent on temperature and/or receiver components, and can be removed by calibration [19.81].

19.7 Receiver Noise and Multipath

Multipath and receiver noise errors have been merged into the error terms $e_{r,j}^s$ and $\epsilon_{r,j}^s$ in (19.6) and (19.9). However, the characteristics of these errors contributions differ significantly. In the following, a brief overview will be provided.

19.7.1 Receiver Noise

Measurement noise is introduced by imperfections of the different electrical components in the signal processing chain, including the antenna, cables, and connectors as well as the receiver itself. Additionally, the antenna receives noise from natural or artificial sources in its environment. This noise introduces random errors

in the pseudorange and carrier-phase observations. The power ratio between this background noise generated by the environment and the GNSS hardware, and a received signal from a navigation satellite, is a measure of the signal strength. A commonly used metric is the carrier-to-noise-power-density ratio C/N_0 , which is the ratio of the power level of the carrier signal C to the noise power N_0 in a 1 Hz bandwidth. Obviously, a large C/N_0 means a strong signal or an ambient noise floor.

The measurement noise of the receiver's DLL for code tracking and its PLL for phase tracking directly depends on the carrier-to-noise-power-density ratio of the received signal (Chap. 14). For $C/N_0 > 35.0$ dB-Hz, the standard deviation of the measurement noise of

a code tracking loop for binary phase shift keying (BPSK)-modulated signals with an early-minus-late correlator can be approximated as

$$\sigma_{\text{DLL}} \approx \sqrt{\frac{dB_L}{2C/N_0}} \lambda_c, \quad (19.49)$$

where d is the correlator spacing in units of code chips, B_L is the equivalent code loop noise bandwidth in Hz, and λ_c is the wavelength of the code. It becomes obvious from this relation that the noise error of the pseudorange observations decreases with increasing C/N_0 . Furthermore, the noise depends on the correlator and on the tracking loop design, as well as on the signal's code chip length [19.82, p. 181]. Current receivers achieve a code measurement noise standard deviation of a decimeter or less when tracking the modernized signals with high chip rates. However, a more sophisticated relation than shown in (19.49) has to be used for modernized signals like the binary offset carrier (BOC) and alternative BOC (AltBOC) signals of Galileo [19.83].

The noise of the carrier-tracking loop can be approximated in a similar manner. The standard deviations of the phase lock loop σ_{PLL} assuming high C/N_0 can be approximated by

$$\sigma_{\text{PLL}} \approx \sqrt{\frac{B_P}{C/N_0}} \frac{\lambda}{2\pi}, \quad (19.50)$$

which depends on the carrier loop noise bandwidth B_P in Hz and on the carrier-phase wavelength λ . Typically, the standard deviation of carrier-phase noise for high C/N_0 is less than a millimeter [19.82, p. 182].

19.7.2 Multipath Errors

Multipath errors in pseudorange and carrier-phase observations originate from the reception of the same

signal through multiple signal paths (Chap. 15). The ground or other reflective surfaces in the vicinity of the antenna can cause reflections of the direct signal. The reflections arrive with a delay, an attenuation, and a phase shift at the antenna, depending on the characteristics of the reflector and on the relative geometry of satellite, reflector, and receiving antenna. The superposition of the direct signal and its reflections is then tracked by the receiver.

In the case of the code tracking loops, the superimposed signal leads to a distortion of the correlation peak, and thus to a range error in the pseudorange measurement. A common misconception is to believe that multipath always causes the measured pseudorange to be longer. Even though the reflected signal always arrives later than the direct signal, it can be shifted in phase, and the correlation function can be distorted such that the receiver tracks *shorter* pseudoranges. If the delay of the signal is larger than the sum of the code chip length and half the correlator spacing, the correlation process is unaffected by multipath errors as long as the direct signal is tracked. Therefore, the multipath characteristics also depend on the width of the correlator spacing, and on the length of the code chip [19.84, 85].

Carrier-phase observations can likewise be affected by multipath errors. The maximum error for phase observations is $\lambda/4$ and it occurs if the reflected signal arrives with the same amplitude as the direct signal, but with a phase shift of 180° [19.86].

Contrary to measurement noise, multipath errors cannot be characterized as random errors. The temporal variation of multipath errors depends on the change in the geometry between satellite, reflector, and receiver. Due to the change in the relative shift of the direct and of the reflected signal, multipath typically has periodical characteristics with timescales of several seconds to several minutes. Furthermore, the errors do not have a mean value of zero and can therefore not be averaged out [19.86].

References

- 19.1 F. van Graas, C. Bartone, T. Arthur: GPS antenna phase and group delay corrections, Proc. ION NTM 2004, San Diego (ION, Virginia 2004) pp. 399–408
- 19.2 T. Murphy, P. Geren, T. Pankaskie: GPS antenna group delay variation induced errors in a GNSS based precision approach and landing systems, Proc. ION GNSS 2007, Fort Worth (ION, Virginia 2007) pp. 2974–2989
- 19.3 P. Axelrad, R.G. Brown: GPS navigation algorithms. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker Jr. (AIAA, Washington 1996) pp. 409–433
- 19.4 P. Enge, P. Misra: *Global Positioning System: Signals, Measurements, and Performance*, 2nd edn. (Ganga-Jamuna Press, Lincoln 2006)
- 19.5 C.-C. Su: Reinterpretation of the Michelson-Morley experiment based on the GPS Sagnac correction, *Europhys. Lett.* **56**(2), 170–174 (2001)
- 19.6 C. Møller: *The Theory of Relativity* (Clarendon Press, Oxford 1952)
- 19.7 S.Y. Zhu, E. Groten: Relativistic Effects in GPS, GPS-Tech. Appl. Geod. Surv. Proc. Int. GPS-Workshop

- 1988, Darmstadt, ed. by E. Groten, R. Strauß (Springer, Berlin, Heidelberg 1988) pp. 41–46
- 19.8 N. Ashby: Relativity in the global positioning system, *Living Rev. Relativ.* **6**(1), 1–42 (2003)
- 19.9 J. Kouba: Relativistic time transformations in GPS, *GPS Solutions* **5**(4), 1–9 (2002)
- 19.10 J. Kouba: Improved relativistic transformations in GPS, *GPS Solutions* **8**(3), 170–180 (2004)
- 19.11 J. Zhang, K. Zhang, R. Grenfell, R. Deakin: Short note: On the relativistic Doppler effect for precise velocity determination using GPS, *J. Geod.* **80**(2), 104–110 (2006)
- 19.12 B. Hoffmann-Wellenhof, H. Lichtenegger, E. Wasle: *GNSS – Global Navigation Satellite Systems* (Springer, Wien, New York 2008)
- 19.13 J.A. Klobuchar: Ionospheric effects on GPS. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington 1996) pp. 485–515
- 19.14 O. Øvstedal: Absolute positioning with single-frequency GPS receivers, *GPS Solutions* **5**(4), 33–44 (2002)
- 19.15 T. Beran, S.B. Bisnath, R.B. Langley: Evaluation of high-precision, single-frequency GPS point positioning models, *Proc. ION GNSS 2004*, Long Beach (ION, Virginia 2004) pp. 1893–1901
- 19.16 H. Zhang, Z. Gao, M. Ge, X. Niu, L. Huang, R. Tu, X. Li: On the convergence of ionospheric constrained precise point positioning (IC-PPP) based on undifferential uncombined raw GNSS observations, *Sensors* **13**(11), 15708–15725 (2003)
- 19.17 T. Beran, D. Kim, R.B. Langley: High-precision single-frequency GPS point positioning, *Proc. ION GPS/GNSS 2003*, Portland (ION, Virginia 2003) pp. 1192–1200
- 19.18 G.D. Thayer: An improved equation for the radio refractive index of air, *Radio Sci.* **9**(10), 803–807 (1974)
- 19.19 J. Saastamoinen: Atmospheric correction for the troposphere and stratosphere in radio ranging satellites. In: *The Use of Artificial Satellites for Geodesy*, ed. by S.W. Henriksen, A. Mancini, B.H. Chovitz (AGU, Washington 1972) pp. 247–251
- 19.20 J.L. Davis, T.A. Herring, I.I. Shapiro, A.E.E. Rogers, G. Elgered: Geodesy by radio interferometry: Effects of atmospheric modeling errors on estimates of baseline length, *Radio Sci.* **20**(6), 1593–1607 (1985)
- 19.21 H.S. Hopfield: Two-quartic tropospheric refractivity profile for correcting satellite data, *J. Geophys. Res.* **74**(18), 4487–4499 (1969)
- 19.22 J. Askne, H. Nordius: Estimation of tropospheric delay for microwaves from surface weather data, *Radio Sci.* **22**(3), 379–386 (1987)
- 19.23 T.A. Herring: Modeling atmospheric delays in the analysis of space geodetic data. In: *Refraction of Atmospheric Signals in Geodesy*, ed. by J.C. Munck, T.A.T. Spoelstra (Netherlands Geodetic Commission, Delft 1992) pp. 157–164
- 19.24 J.W. Marini: Correction of satellite tracking data for an arbitrary tropospheric profile, *Radio Sci.* **7**(2), 223–231 (1999)
- 19.25 A.E. Niell: Global mapping functions for the atmosphere delay at radio wavelengths, *J. Geophys. Res.* **101**(B2), 3227–3246 (1996)
- 19.26 A.E. Niell: Improved atmospheric mapping functions for VLBI and GPS, *Earth Planets Space* **52**(10), 699–702 (2000)
- 19.27 J. Böhm, B. Werl, H. Schuh: Troposphere mapping functions for GPS and very long baseline interferometry from European Centre for Medium-Range Weather Forecasts operational analysis data, *J. Geophys. Res. Solid Earth* (1978–2012) **111**(B2), 1–9 (2006)
- 19.28 P. Collins, R.B. Langley, J. LaMance: Limiting factors in tropospheric propagation delay error modelling for GPS airborne navigation, *Proc. ION AM 1996*, Cambridge (ION, Virginia 1996) pp. 519–528
- 19.29 R.F. Leandro, M.C. Santos, R.B. Langley: UNB neutral atmosphere models: Development and performance, *Proc. ION NTM 2006*, Monterey (ION, Virginia 2006) pp. 564–573
- 19.30 J. Böhm, R. Heinkelmann, H. Schuh: Short note: A global model of pressure and temperature for geodetic applications, *J. Geod.* **81**(10), 679–683 (2007)
- 19.31 J. Böhm, A. Niell, P. Tregoning, H. Schuh: Global mapping function (GMF): A new empirical mapping function based on numerical weather model data, *Geophys. Res. Lett.* **33**(L07304), 1–4 (2006)
- 19.32 A.K. Tetewsky, F.E. Mullen: Carrier phase wrap-up induced by rotating GPS antennas, *Proc. ION AM 1996*, Cambridge (ION, Virginia 1996) pp. 21–28
- 19.33 M. Garcia-Fernández, M. Markgraf, O. Montenbruck: Spin rate estimation of sounding rockets using GPS wind-up, *GPS Solutions* **12**(3), 155–161 (2008)
- 19.34 J.T. Wu, S.C. Wu, G.A. Hajj, W.I. Bertiger, S.M. Lichten: Effects of antenna orientation on GPS carrier phase, *Manuscr. Geod.* **18**(2), 91–98 (1993)
- 19.35 Y.E. Bar-Sever: A new model for GPS yaw attitude, *J. Geod.* **70**(11), 714–723 (1996)
- 19.36 J. Kouba: A simplified yaw-attitude model for eclipsing GPS satellites, *GPS Solutions* **13**(1), 1–12 (2009)
- 19.37 F. Dilssner: GPS IIF-1 satellite, antenna phase center and attitude modeling, *Inside GNSS* **5**(6), 59–64 (2010)
- 19.38 F. Dilssner, T. Springer, W. Enderle: GPS IIF yaw attitude control during eclipse season, *Proc. AGU Fall Meet.*, San Francisco (AGU, Washington 2011) pp. 1–23
- 19.39 F. Dilssner, T. Springer, G. Gienger, J. Dow: The GLONASS-M satellite yaw-attitude model, *Adv. Space Res.* **47**(1), 160–171 (2010)
- 19.40 P. Zentgraf, H.-D. Fischer, L. Kaffer, A. Konrad, E. Lehl, C. Müller, W. Oesterlin, M. Wiegand: AOC design and test for GSTB-V2B, *Proc. 6th Int. ESA Conf. Guid. Navig. Contr. Syst.*, Loutraki, ed. by D. Danesy (ESA, Noordwijk 2005) pp. 1–7
- 19.41 A. Konrad, H.-D. Fischer, C. Müller, W. Oesterlin: Attitude & orbit control system for Galileo IOV, *Proc. 17th IFAC Symp. Autom. Contr. Aerosp.*, Toulouse, ed. by H. Siguerdjane (IFAC, Laxenburg 2007) pp. 25–30

- 19.42 Y. Ishijima, N. Inaba, A. Matsumoto, K. Terada, H. Yonechi, H. Ebisutani, S. Ukawa, T. Okamoto: Design and development of the first quasi-zenith satellite attitude and orbit control system, *IEEE Aerosp. Conf.*, Big Sky (2009)
- 19.43 A. Hauschild, P. Steigenberger, C. Rodriguez-Solano: QZS-1 yaw attitude estimation based on measurements from the CONGO network, *Navigation* **59**(3), 237–248 (2012)
- 19.44 S. Zhou, X. Hu, J. Zhou, J. Chen, X. Gong, C. Tang, B. Wu, L. Liu, R. Guo, F. He, X. Li, H. Tan: Accuracy analyses of precise orbit determination and timing for COMPASS/Beidou-2 4GEO/5IGSO/4MEO constellation, *Proc. CSNC 2013*, Vol. III, Wuhan, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin 2013) pp. 89–102
- 19.45 W. Wang, G. Chen, S. Guo, X. Song, Q. Zhao: A study on the Beidou IGSO/MEO satellite orbit determination and prediction of the different yaw control mode, *China Satell. Navig. Conf. (CSNC)*, Wuhan, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin Heidelberg 2013) pp. 31–40
- 19.46 J. Guo, Q. Zhao, T. Geng, X. Su, J. Liu: Precise orbit determination for COMPASS IGSO satellites during yaw maneuvers, *Proc. CSNC 2013*, Wuhan, Vol. III, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin 2013) pp. 41–53
- 19.47 T. Kersten, S. Schön: GNSS group delay variations – Potential for improving GNSS based time and frequency transfer?, *Proc. 43rd Annu. PTI Syst. Appl. Meet.*, Long Beach (ION, Virginia 2011) pp. 255–270
- 19.48 B. Görres, J. Campbell, M. Becker, M. Siemes: Absolute calibration of GPS antennas: Laboratory results and comparison with field and robot techniques, *GPS Solutions* **10**(2), 136–145 (2006)
- 19.49 M. Schmitz, G. Wübbena, G. Boettcher: Tests of phase center variations of various GPS antennas, and some results, *GPS Solutions* **6**(1/2), 18–27 (2002)
- 19.50 F. Menge, G. Seeber, C. Völksen, G. Wübbena, M. Schmitz: Results of absolute field calibration of GPS antenna PCV, *Proc. ION GPS 1998*, Nashville (ION, Virginia 1998) pp. 31–38
- 19.51 G.L. Mader: GPS antenna calibration at the National Geodetic Survey, *GPS Solutions* **3**(1), 50–58 (1999)
- 19.52 M. Rothacher: Comparison of absolute and relative antenna phase center variations, *GPS Solutions* **4**(4), 55–60 (2001)
- 19.53 R. Schmid, M. Rothacher: Estimation of elevation-dependent satellite antenna phase center variations of GPS satellites, *J. Geod.* **77**(7), 440–446 (2003)
- 19.54 S.Y. Zhu, F.-H. Massmann, Y. Yu, C. Reigber: Satellite antenna phase center offsets and scale errors in GPS solutions, *J. Geod.* **76**(11/12), 668–672 (2003)
- 19.55 B.R. Schupler: Signal characteristics of GPS user antennas, *Navigation* **41**(3), 277–296 (1994)
- 19.56 G. Wübbena, M. Schmitz, F. Menge, G. Seeber, C. Völksen: A new approach for field calibration of absolute antenna phase center variations, *Navigation* **44**(2), 247–256 (1997)
- 19.57 G. Seeber, F. Menge, C. Völksen, G. Wübbena, M. Schmitz: Precise GPS positioning improvements by reducing antenna and site dependent effects. In: *Advances in Positioning and Reference Frames*, ed. by F.K. Brunner (Springer, Berlin 1998) pp. 237–244
- 19.58 G. Wübbena, M. Schmitz, F. Menge, V. Böder, G. Seeber: Automated absolute field calibration of GPS antennas in real-time, *Proc. ION GPS 2000*, Salt Lake City (ION, Virginia 2000) pp. 2512–2522
- 19.59 G. Wübbena, M. Schmitz, G. Mader, F. Czapke: GPS Block II/IIA satellite antenna testing using the automated absolute field calibration with robot, *Proc. ION GNSS 2007*, Fort Worth (ION, Virginia 2007) pp. 1236–1243
- 19.60 R. Schmid, P. Steigenberger, G. Gendt, M. Ge, M. Rothacher: Generation of a consistent absolute phase center correction model for GPS receiver and satellite antennas, *J. Geod.* **81**(12), 781–798 (2007)
- 19.61 M. Rothacher, R. Schmid: ANTEX: The Antenna Exchange Format, Version 1.4 (2010) <http://www.igs.org/assets/txt/antex14.tx>
- 19.62 C. Hegarty, E. Powers, B. Fonville: Accounting for timing biases between GPS, modernized GPS, and Galileo signals, *Proc. 36th Annu. PTI Meet.*, Washington (PTI, Washington 2004) pp. 307–317
- 19.63 O. Montenbruck, A. Hauschild: Code biases in multi-GNSS point positioning, *Proc. 2013 Int. Tech. Meet. Inst. Navig.*, San Diego (ION, Virginia 2013) pp. 616–628
- 19.64 D. Odijk, P.J.G. Teunissen: Characterization of between-receiver GPS-Galileo inter-system biases and their effect on mixed ambiguity resolution, *GPS Solutions* **17**(4), 521–533 (2013)
- 19.65 J. Paziewski, P. Wielgosz: Accounting for Galileo-GPS inter-system biases in precise satellite positioning, *J. Geod.* **89**(1), 81–93 (2015)
- 19.66 A. Dalla Torre, A. Caporali: An analysis of intersystem biases for multi-GNSS positioning, *GPS Solutions* **19**(2), 297–307 (2015)
- 19.67 D. Kozlov, M. Tkachenko: Centimeter-level real-time kinematic positioning with GPS+GLONASS C/A receivers, *Navigation* **45**(2), 137–147 (1998)
- 19.68 G. Bishop, A. Mazella, E. Holland, S. Rao: Algorithms that use the ionosphere to control GPS errors, *Proc. IEEE PLANS 1996*, Atlanta (1996)
- 19.69 R. Warnant: Reliability of the TEC computed using GPS measurements – The problem of hardware biases, *Acta Geod. Geophys. Hung.* **32**(3/4), 451–459 (1997)
- 19.70 O. Montenbruck, U. Hugentobler, R. Dach, P. Steigenberger, A. Hauschild: Apparent clock variations of the Block IIF-1 (SVN62) GPS satellite, *GPS Solutions* **16**(3), 303–313 (2012)
- 19.71 Navstar GPS Space Segment/Navigation User Segment Interfaces, Interface Specification (Global Positioning Systems Directorate, California 2013) IS-GPS-200H, 24 Sep. 2013
- 19.72 A. Hauschild, O. Montenbruck: A study on the dependency of GNSS pseudorange biases on correlator spacing, *GPS Solutions* **20**(2), 159–171 (2016)
- 19.73 A. Hauschild, O. Montenbruck: The Effect of correlator and front-end design on GNSS pseudorange biases for geodetic receivers, *Proc. ION GNSS+ 2015*,

- Tampa (ION, Virginia 2015) pp. 2835–2844
- 19.74 M. Ge, G. Gendt, M. Rothacher, C. Shi, J. Liu: GPS carrier-phase ambiguities in precise point positioning (PPP) with daily observations, *J. Geod.* **82**(7), 389–399 (2008)
- 19.75 J. Geng, F.N. Teferle, C. Shi, X. Meng, A.H. Dodson, J. Liu: Integer ambiguity resolution in precise point positioning with hourly data, *GPS Solutions* **13**(4), 263–270 (2009)
- 19.76 D. Laurichesse, F. Mercier, J.P. Berthias: Zero-difference integer ambiguity fixing on single frequency receivers, *Proc. ION GNSS 2009, Savannah* (ION, Virginia 2009) pp. 2460–2469
- 19.77 D. Laurichesse, F. Mercier, J.P. Berthias, P. Broca, L. Cerri: Integer ambiguity resolution undifferenced GPS phase measurements and its application to PPP and satellite precise orbit determination, *Navigation* **56**(2), 135–149 (2009)
- 19.78 P. Collins, S. Bisnath, F. Lahaye, P. Heroux: Undifferenced GPS ambiguity resolution using the decoupled clock model and ambiguity datum fixing, *Navigation* **57**(2), 123–135 (2010)
- 19.79 P.J.G. Teunissen, A. Khodabandeh: Review and principles of PPP-RTK methods, *J. Geod.* **89**(3), 217–240 (2015)
- 19.80 L. Wanninger: Carrier-phase inter-frequency biases of GLONASS receivers, *J. Geod.* **86**(2), 139–148 (2012)
- 19.81 J.-M. Sleewaegen, A. Simsky, W. de Wilde, F. Boon, T. Willems: Origin and compensation of GLONASS inter-frequency carrier phase biases in GNSS receivers, *Proc. ION GNSS 2012, Nashville* (ION, Virginia 2012) pp. 2995–3001
- 19.82 P.J.G. Teunissen, A. Kleusberg (Eds.): *GPS for Geodesy*, 2nd edn. (Springer, Berlin, Heidelberg, New York 1998)
- 19.83 J.-M. Sleewaegen, W. de Wilde, M. Hollreiser: Galileo AltBOC receiver, *Proc. ENC-GNSS 2004, Rotterdam* (Netherlands Institute of Navigation, Rotterdam 2004) pp. 1–9
- 19.84 M.S. Braasch: Multipath effects. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker Jr. (AIAA, Washington 1996) pp. 547–568
- 19.85 G. Seeber: *Satellite Geodesy*, 2nd edn. (Walter de Gruyter, Berlin, New York 2003)
- 19.86 M. Irsigler: Multipath Propagation, Mitigation and Monitoring in the Light of Galileo and the Modernized GPS, Ph.D. Thesis (TU München, München 2008)

Combination

20. Combinations of Observations

André Hauschild

Part D | 20.1

This chapter introduces the concept of observation combinations, commonly used, for example, to compute positioning solutions with measurements from multiple frequencies or to study measurement noise, multipath, or ionospheric effects. Based on a generic parametrization for pseudorange and carrier-phase observations, a general expression for linear combinations is introduced. The impact of the coefficients on the properties and the noise of the combined observable is explained. The chapter covers combinations using measurements from a single satellite observed by one receiver. The discussion will then be extended to differential observations from two satellites, receivers and epochs.

20.1	Fundamental Equations	583
20.2	Combinations of Single-Satellite and Single-Receiver Observations	586
20.2.1	Narrow- and Wide-Lane Combinations	586
20.2.2	Ionosphere Combination	589
20.2.3	Ionosphere-Free Combination	590
20.2.4	Multipath Combination	591
20.3	Combinations of Multisatellite and Multireceiver Observations	594
20.3.1	Between-Receiver Single Difference	594
20.3.2	Between-Satellite Single Difference	596
20.3.3	Double Difference	596
20.3.4	Triple Difference	598
20.3.5	Single and Double Difference on Zero-Baselines	599
20.4	Pseudorange Filtering	601
	References	603

20.1 Fundamental Equations

Forming combinations from multiple observations of the same or different types can be useful for various aspects of global navigation satellite system (GNSS) data processing or analysis. In particular, combining observations can conveniently be used to eliminate nuisance parameters, like, for example, the ionospheric delay. Furthermore, they are a useful measure to isolate and emphasize particular errors or features to be studied, like, for example, multipath errors or ambiguity parameters.

For ease of notation when deriving the combination equations, it is helpful to introduce a term $P_{r,IF}^s$ [20.1], which is defined as

$$P_{r,IF}^s(t) = \rho_r^s(t) + c \left(dt_r(t) - dt^s(t) + \delta t^{\text{rel}}(t) \right) + T_r^s(t). \quad (20.1)$$

In the notation adopted for this chapter, the superscript s denotes the observed satellite and the subscript r identifies the receiver. The term $P_{r,IF}^s$ only depends on the geometric range ρ_r^s , the receiver clock offset dt_r , the satellite clock offset dt^s , the relativistic term

δt^{rel} , and the tropospheric delay T_r^s . All timing delays are converted to units of length by multiplication with the speed of light c . Note that not only the ionospheric delay $I_{r,j}^s$, but also all other frequency-dependent terms like receiver bias $d_{r,j}$ and satellite bias d_j^s , line-of-sight-dependent group-delay variations (or code-phase pattern) $\xi_{r,j}^s$, as well as multipath and noise $e_{r,j}^s$ are excluded. The subscript j is used here as frequency or signal identifier. The full pseudorange observation equation as in (19.6) of Chap. 19 can be restored from

$$p_{r,j}^s(t) = P_{r,IF}^s(t) + I_{r,j}^s(t) + \xi_{r,j}^s(t) + cd_{r,j}^s + e_{r,j}^s(t). \quad (20.2)$$

The reader may refer to Chap. 19 for a more detailed discussion of the individual terms. An abbreviated notation has been used in (20.2), which combines the receiver bias $d_{r,j}$ and satellites bias d_j^s as follows

$$d_{r,j}^s = d_{r,j} + d_j^s. \quad (20.3)$$

When combinations of observations from satellites of different constellations are formed, a property called inter-system bias (ISB) becomes relevant. The ISB is

generally understood as a combination of delays (or biases) and a system-time offset between two constellations. It has not explicitly been noted in (20.2), since it only manifests itself on differences of observations. However, the contribution of the system-time offset to the ISB is implicitly included in the satellite clock offset term in (20.1) and the contribution of the delays in the signal reception and processing chain is contained in the bias term in (20.2). It is therefore important to note that the clock offsets in (20.1) refer to a single, common time system for all constellations. The clock parameters, which are transmitted in the satellites' navigation message or provided as external corrections, typically refer to the system-time of the corresponding GNSS, however. The subtlety of the ISB will further be elaborated in the discussion of single- and double differences of observations.

An equivalent expression to (20.2) can also be found for the carrier-phase observation $\varphi_{r,j}^s$ ((19.9) in Chap. 19)

$$\begin{aligned} \varphi_{r,j}^s(t) = & P_{r,IF}^s(t) - I_{r,j}^s(t) + \zeta_{r,j}^s(t) + c\delta_{r,j}^s \\ & + \lambda_j (\omega_r^s(t) + N_{r,j}^s) + \epsilon_{r,j}^s(t) . \end{aligned} \quad (20.4)$$

In case of the carrier-phase, the correction term due to antenna phase center offset and variation $\zeta_{r,j}^s$, the combined receiver and satellite bias $\delta_{r,j}^s$, the phase wind-up correction ω_r^s , the integer ambiguity $N_{r,j}^s$, and the receiver noise and multipath $\epsilon_{r,j}^s$ appear as extra terms. Note that ω_r^s and $N_{r,j}^s$ have been converted from cycles to units of length by multiplication with the wavelength λ_j . Similar to the pseudorange observations, differences of carrier-phase observations between satellites of different constellations will be affected by an ISB, which is again implicitly included in the satellite clock offset term and the carrier-phase biases. Division of (20.4) by the corresponding wavelength yields an expression for the carrier-phase in units of cycles

$$\Phi_{r,j}^s(t) = \frac{\varphi_{r,j}^s(t)}{\lambda_j} . \quad (20.5)$$

For later developments, it should also be remembered that the first-order ionospheric delay depends on the frequency f_j and on the total electron content (TEC) along the signal propagation path from the receiver to the satellite as follows ((19.29) in Chap. 19)

$$I_{r,j}^s = \frac{40.3 \text{ TEC}}{f_j^2} = \frac{f_1^2}{f_j^2} I_{r,1}^s . \quad (20.6)$$

In (20.6), f_1 is an arbitrarily selected reference frequency and $I_{r,1}^s$ is the corresponding ionospheric delay.

Higher order ionospheric delays are neglected for simplicity.

Generalized equations for linear combinations of GNSS pseudorange and carrier-phase observations have been introduced in [20.2] and are presented here only slightly modified. Assuming that n different signals are available, the combined observable o_c is the linear combinations of carrier-phase observations φ_j , scaled with the coefficient α_j , and pseudorange observations p_j , multiplied with the factor β_j

$$o_{r,c}^s(t) = \sum_{j=1}^n (\alpha_j \varphi_{r,j}^s(t) + \beta_j p_{r,j}^s(t)) . \quad (20.7)$$

Note that at this point any real number is allowed as the value of coefficients α_j and β_j . Therefore, (20.7) also allows to form code- or phase-only combinations or to use a different number of code and phase observations from different frequencies. Substituting (20.2)–(20.4), and (20.6) into (20.7) yields

$$\begin{aligned} o_{r,c}^s(t) = & \left(\sum_{j=1}^n (\alpha_j + \beta_j) \right) P_{r,IF}^s(t) \\ & - \left(\sum_{j=1}^n (\alpha_j - \beta_j) \frac{f_1^2}{f_j^2} \right) I_{r,1}^s(t) \\ & + \left(\sum_{j=1}^n (\alpha_j \zeta_{r,j}^s(t) + \beta_j \epsilon_{r,j}^s(t)) \right) \\ & + \left(\sum_{j=1}^n (\alpha_j \delta_{r,j}^s + \beta_j d_{r,j}^s) \right) c \\ & + \left(\sum_{j=1}^n \alpha_j \lambda_j N_{r,j}^s \right) + \left(\sum_{j=1}^n \alpha_j \lambda_j \right) \omega_r^s(t) \\ & + \left(\sum_{j=1}^n (\alpha_j \epsilon_{r,j}^s(t) + \beta_j e_{r,j}^s(t)) \right) . \end{aligned} \quad (20.8)$$

It becomes obvious from (20.8) that depending on the selection of the coefficient α_j and β_j some terms in the equations can be retained, attenuated or completely eliminated. For example, the magnitude of the geometric range, clock offsets, and tropospheric delay contained in $P_{r,IF}^s(t)$ is controlled through

$$\sum_{j=1}^n (\alpha_j + \beta_j) = h_1 . \quad (20.9)$$

The scaling factor h_1 can take any value depending on the coefficients; however, there are two cases which are commonly used and deserve special attention: if the coefficients are selected such that $h_1 = 1$, the combination is referred to as *geometry-preserving*. If $h_1 = 0$, the term $P_{r,IF}^s(t)$ is completely eliminated, and the combination is *geometry-free*.

The first-order ionospheric delay in (20.8) can be scaled through the factor

$$-\sum_{j=1}^n (\alpha_j - \beta_j) \frac{f_1^2}{f_j^2} = h_2. \quad (20.10)$$

Selecting the coefficients such that $h_2 = 0$ creates a *ionosphere-free* combination. It should be noted that observation combinations can be ionosphere-free and either geometry-free or geometry-preserving at the same time.

If carrier-phase observations are used in the combined observation (20.8), it is useful to introduce a combined ambiguity N_c and the corresponding wavelength λ_c . The sum of the individual ambiguities in (20.8) is related to the combined ambiguity as follows

$$\lambda_c N_c = \sum_{j=1}^n \alpha_j \lambda_j N_j = \lambda_c \sum_{j=1}^n i_j N_j, \quad (20.11)$$

where the integer-phase coefficient i_j has been newly introduced as

$$i_j = \alpha_j \frac{\lambda_j}{\lambda_c} = \alpha_j \frac{f_c}{f_j}, \quad (20.12)$$

and the combined ambiguity is

$$N_c = \sum_{j=1}^n i_j N_j. \quad (20.13)$$

Since all i_j are assumed to be integers, the resulting combined ambiguity N_c is also integer. The combined frequency f_c and wavelength $\lambda_c = c/f_c$ can be found from the sum of all coefficients α as

$$f_c = \frac{\sum_{j=1}^n i_j f_j}{\sum_{j=1}^n \alpha_j} \quad (20.14)$$

and

$$\lambda_c = \frac{\sum_{j=1}^n \alpha_j}{\sum_{j=1}^n i_j \frac{f_j}{c}} = \frac{\sum_{j=1}^n \alpha_j}{\sum_{j=1}^n \frac{i_j}{\lambda_j}}. \quad (20.15)$$

The stochastic properties of the combined observation are also governed by the selection of coefficients in (20.8). Assume that a vector of uncombined observations \mathbf{o} is transformed via a transformation matrix \mathbf{T} into a vector of combined observations \mathbf{o}_c as follows

$$\mathbf{o}_c = \mathbf{T} \mathbf{o}. \quad (20.16)$$

The transformation matrix \mathbf{T} contains the coefficients α and β for the different measurement combinations. The covariance matrix $\mathbf{Q}_{o_c o_c}$ of the combined observations is then found from the covariance matrix \mathbf{Q} of the uncombined observations

$$\mathbf{Q}_{o_c o_c} = \mathbf{T} \mathbf{Q}_{oo} \mathbf{T}^T. \quad (20.17)$$

Assume that the individual observations in \mathbf{o} are uncorrelated and combined into a single observable o_c . With the standard deviation $\sigma_{e,j}$ for pseudorange and $\sigma_{e,j}$ for carrier-phase, (20.17) simplifies to the expression

$$\sigma_c = \sqrt{\sum_{j=1}^n (\alpha_j^2 \sigma_{e,j}^2 + \beta_j^2 \sigma_{e,j}^2)} \quad (20.18)$$

for the standard deviation of the combined observable. In a similar way, the covariances between the combined observations can be obtained. Both the variances and covariances are needed to properly weigh the combined observations in further processing.

The following sections will introduce observation combinations of pseudorange and carrier-phase measurements. Doppler measurements are not considered here. However, for some applications Doppler combinations have practical use, especially in case of the single-, double-, and triple differences or the ionosphere-free combination. In these cases, the Doppler combinations are formed equivalently to the pseudorange and carrier-phase observations.

20.2 Combinations of Single-Satellite and Single-Receiver Observations

In this section, observation combinations are introduced, which use measurements from one receiver and one satellite on one or more frequencies. The combinations have been categorized according to the application they have been designed for, namely:

- Carrier-phase ambiguity resolution
- Isolation or elimination of ionospheric errors
- Multipath analysis.

A multitude of combinations can already be found for each of these categories with only the dual-frequency legacy signals. Even more options are available with the additional frequencies of the modernized GNSSs. Therefore, the derivations have been confined to the most commonly used combinations, and only selected examples with modernized GNSS signals are presented, which have so far been found to be of practical interest.

20.2.1 Narrow- and Wide-Lane Combinations

It is obvious from (20.14) and (20.15) that combining two or more carrier-phase observations into a new signal leads to a different frequency and wavelength for this combination. In case of a combination for which $\sum \alpha = 1$, the combined frequency f_c is

$$f_c = \sum_{j=1}^n i_j f_j. \quad (20.19)$$

Noting that all frequencies of a GNSS are derived from a single base frequency f_0 by multiplication with an integer k_j , the individual frequency is obtained from $f_j = k_j f_0$. Substituting this expression into (20.19) yields

$$f_c = \left(\sum_{j=1}^n i_j k_j \right) f_0 = k f_0, \quad (20.20)$$

where k is called the lane number. The corresponding wavelength is

$$\lambda_c = \frac{c}{k f_0} = \frac{\lambda_0}{k}, \quad (20.21)$$

where λ_0 is the wavelength of the base frequency f_0 . Since all i_j and k_j are integers, k is also an integer. This parameter uniquely defines the frequency and wavelength of the new signal combination [20.3]. For $k = 1$, the combined wavelength is equal to λ_0 , which is the

longest wavelength that can be achieved. For the Global Positioning System (GPS) and the Quasi-Zenith Satellite System (QZSS), the base frequency is 10.230 MHz with a corresponding wavelength of ≈ 29.31 m. In case of the BeiDou Phase II signals as listed in Table 20.1, the base frequency f_0 is 2.046 MHz and the largest possible wavelength is ≈ 146.53 m [20.4]. By selecting larger values for k , combined signals with different wavelengths can be created, which are all fractions of λ_0 .

With the help of the lane number k , the combinations can be categorized into three groups [20.3]:

1. The *wide-lane* combinations, for which the combined wavelength is larger than the largest individual wavelength in the combination.
2. The *intermediate-lane* combinations, for which λ_c lies between the largest and the shortest individual wavelength.
3. The *narrow-lane* combinations, which have a shorter wavelength than the individual signal with the shortest wavelength in the combination.

The group of wide-lane combinations is especially interesting for the purpose of integer ambiguity resolution (Chap. 23) due to their long wavelength. In that case the wide-lane combinations are applied to the double-differenced observations. With observations on multiple frequencies available, many different combinations not only for wide-lanes but for each of the categories can be formed. Even with only dual-frequency measurements, there is a range of possibilities to form wide-lane combinations which have different characteristics concerning the reduction of measurement noise and ionospheric delay [20.5, 6].

In the following, the most common dual-frequency wide-lane (WL) combination will be introduced. It is designed to be a geometry-preserving combination using only carrier-phase measurements on the frequencies f_A and f_B . By selecting the integer coefficients as $i_A = +1$ and $i_B = -1$, one obtains the WL carrier-phase combination $\Phi_{r,WL}^s$ in units of cycles

$$\Phi_{r,WL}^s = \Phi_{r,A}^s - \Phi_{r,B}^s = \frac{\varphi_{r,A}^s}{\lambda_A} - \frac{\varphi_{r,B}^s}{\lambda_B}. \quad (20.22)$$

The corresponding wavelength λ_{WL} is

$$\lambda_{WL} = \frac{c}{f_A - f_B}. \quad (20.23)$$

Multiplication of (20.22) with (20.23) leads to the carrier-phase wide-lane combination $\varphi_{r,WL}^s$ in units of

meters

$$\varphi_{r,WL}^s = \frac{f_A}{f_A - f_B} \varphi_{r,A}^s - \frac{f_B}{f_A - f_B} \varphi_{r,B}^s. \quad (20.24)$$

In a similar way, a geometry-preserving narrow-lane carrier-phase combination can be formed using the integer coefficients $i_A = +1$ and $i_B = +1$

$$\varphi_{r,NL}^s = \varphi_{r,A}^s + \varphi_{r,B}^s = \frac{\varphi_{r,A}^s}{\lambda_A} + \frac{\varphi_{r,B}^s}{\lambda_B}, \quad (20.25)$$

with the corresponding NL wavelength

$$\lambda_{NL} = \frac{c}{f_A + f_B}. \quad (20.26)$$

The narrow-lane combination in units of meters $\varphi_{r,NL}^s$ is then

$$\varphi_{r,NL}^s = \frac{f_A}{f_A + f_B} \varphi_{r,A}^s + \frac{f_B}{f_A + f_B} \varphi_{r,B}^s. \quad (20.27)$$

Evaluating the expressions for the wavelengths in (20.23) and (20.26) illustrates why these combinations bear their names: using, for example, the heritage GPS signals on the frequencies L1 and L2, the wavelength of the narrow-lane combination yields ≈ 10.7 cm. Obviously, the narrow-lane observable has a decreased combined wavelength compared to the original observations. The wide-lane combination, on the contrary, has a wavelength of ≈ 0.86 m for GPS L1/L2.

In case of modernized GPS with signals from three frequencies available, the carrier-phase observations on L1 and L5 can be used in (20.20) to form a combination with a wavelength of ≈ 0.75 m, which is sometimes referred to as *medium-lane* (ML) signal. Using the signals on the two closer frequencies L2 and L5 in (20.22) yields a wavelength of ≈ 5.86 m, typically referred to as *extra wide-lane* (EWL) [20.7]. Similar combinations are also possible for other GNSSs, which transmit on more than two frequencies. The three-carrier ambiguity resolution (TCAR) technique for Galileo [20.8, 9] and the cascade integer resolution (CIR) technique for GPS [20.10, 11] utilize these wide-lane and extra wide-lane combinations. Table 20.1 lists the resulting wavelengths of dual-frequency wide-lane and narrow-lane combinations for different GNSS signals. Wide laning is often a good starting point in the construction of the decorrelating Z-transformation (Chap. 23). For a comparison of TCAR, CIR, and the least-squares ambiguity decorrelation adjustment (LAMBDA), see [20.12–14].

Evaluation of the combination factors in (20.24) and (20.27) with (20.18) yields the standard deviation

of the measurement noise for the wide-lane combination σ_{WL} and for the narrow-lane combination σ_{NL} in units of meters

$$\sigma_{NL} = \sqrt{\frac{f_A^2}{(f_A + f_B)^2} \sigma_A^2 + \frac{f_B^2}{(f_A + f_B)^2} \sigma_B^2}, \quad (20.28)$$

$$\sigma_{WL} = \sqrt{\frac{f_A^2}{(f_A - f_B)^2} \sigma_A^2 + \frac{f_B^2}{(f_A - f_B)^2} \sigma_B^2}. \quad (20.29)$$

Assuming identical noise σ for the carrier-phase signals involved in the combinations, one obtains

$$\sigma_{NL} = \sqrt{\frac{f_A^2 + f_B^2}{(f_A + f_B)^2}} \sigma \quad (20.30)$$

and

$$\sigma_{WL} = \sqrt{\frac{f_A^2 + f_B^2}{(f_A - f_B)^2}} \sigma. \quad (20.31)$$

Evaluating these expressions for the GPS frequencies, L1 and L2 shows that the noise of the narrow-lane combination is reduced by a factor of ≈ 0.71 compared to that of an individual carrier-phase measurement. It is interesting to note that this factor does not only apply for GPS L1 and L2, but is also approximately the same for all current GNSS signal combinations of observations in the lower (1100–1300 MHz) and upper L-band (near 1600 MHz). Forming the wide-lane combination with GPS L1 and L2, on the other hand, leads to a significantly increased noise, which is ≈ 5.7 times higher than the noise of the individual observations. If adjacent frequencies are used, the increase in noise can be even higher, for example choosing GPS L2 and L5 leads to a factor of ≈ 33.2 , using Galileo E5a and E5b yields a factor of ≈ 65.8 . Though the increase in noise becomes large when the difference of the frequencies in the wide-lane combination is small, it is obvious from Table 20.1 that the resulting carrier-phase noise is still small compared to the increase in the corresponding wavelength.

Of course, narrow-lane and wide-lane combinations can also be formed from pseudoranges by using these observations in place of the carrier-phase in (20.24) and (20.27). This is exploited in the Melbourne–Wübbena combination o_{MW} , which is formed from the difference of narrow-lane pseudorange observations and wide-lane carrier-phase observations [20.15, 16]

$$\begin{aligned} o_{MW} &= \varphi_{r,WL}^s - p_{r,NL}^s \\ &= \frac{f_A}{f_A - f_B} \varphi_{r,A}^s - \frac{f_B}{f_A - f_B} \varphi_{r,B}^s \\ &\quad - \frac{f_A}{f_A + f_B} p_{r,A}^s - \frac{f_B}{f_A + f_B} p_{r,B}^s. \end{aligned} \quad (20.32)$$

Table 20.1 Wavelength of wide-lane combination (20.24) in meters (upper right triangular part, underlined) and wavelength of narrow-lane combination (20.27) in centimeters (lower left triangular part) for different dual-frequency GNSS signals. Base frequency f_0 and integer multipliers k_j are listed in parentheses in the header line for code division multiple access (CDMA)-based constellations. Individual carrier-phase frequencies f_j are listed in parentheses in first column. All frequencies are in MHz

GPS ($f_0 = 10.230$)	L1 (154)		L2 (120)		L5 (115)
L1 (1575.420)	–		<u>0.86</u>		<u>0.75</u>
L2 (1227.600)	10.7		–		<u>5.86</u>
L5 (1176.450)	10.9		12.5		–
QZSS ($f_0 = 10.230$)	L1 (154)	LEX (125)	L2 (120)		L5 (115)
L1 (1575.420)	–	<u>1.01</u>	0.86		0.75
LEX (1278.750)	10.5	–	<u>5.90</u>		<u>2.94</u>
L2 (1227.600)	10.7	12.0	–		<u>5.86</u>
L5 (1176.450)	10.9	12.2	12.5		–
GALILEO ($f_0 = 5.115$)	E1 (308)	E6 (250)	E5b (236)	E5 (233)	E5a (230)
E1 (1575.420)	–	<u>1.01</u>	<u>0.81</u>	<u>0.78</u>	<u>0.75</u>
E6 (1278.750)	10.5	–	<u>4.19</u>	<u>3.45</u>	<u>2.93</u>
E5b (1207.140)	10.8	12.1	–	–	<u>9.77</u>
E5 (1191.795)	10.8	12.1	–	–	–
E5a (1176.450)	10.9	12.2	12.6	–	–
GLONASS	L1	L2	L3		
L1 ^a (1602.000)	–	<u>0.84</u>	<u>0.75</u>		
L2 ^a (1246.000)	10.5	–	<u>6.82</u>		
L3 ^b (1202.025)	10.7	12.2	–		
BEIDOU ^c ($f_0 = 2.046$)	B1 (763)	B3 (620)	B2 (590)		
B1 (1561.098)	–	<u>1.02</u>	<u>0.85</u>		
B3 (1268.520)	10.6	–	<u>4.88</u>		
B2 (1207.140)	10.8	12.1	–		

^a FDMA frequency for channel number 0

^b CDMA frequency for GLONASS-K1 satellites

^c BeiDou Phase II signals

This combination fulfills both conditions for geometry-free and ionosphere-free combinations. Substituting (20.2) and (20.24) into (20.32) yields

$$\begin{aligned}
 o_{\text{MW}} = & \lambda_{\text{WL}} (N_{\text{r,A}}^{\text{s}} - N_{\text{r,B}}^{\text{s}}) \\
 & + \frac{f_{\text{A}}}{f_{\text{A}} - f_{\text{B}}} c \delta_{\text{r,A}}^{\text{s}} - \frac{f_{\text{B}}}{f_{\text{A}} - f_{\text{B}}} c \delta_{\text{r,B}}^{\text{s}} \\
 & - \frac{f_{\text{A}}}{f_{\text{A}} + f_{\text{B}}} c d_{\text{r,A}}^{\text{s}} - \frac{f_{\text{B}}}{f_{\text{A}} + f_{\text{B}}} c d_{\text{r,B}}^{\text{s}} \\
 & + \frac{f_{\text{A}}}{f_{\text{A}} - f_{\text{B}}} \zeta_{\text{r,A}}^{\text{s}} - \frac{f_{\text{B}}}{f_{\text{A}} - f_{\text{B}}} \zeta_{\text{r,B}}^{\text{s}} \\
 & - \frac{f_{\text{A}}}{f_{\text{A}} + f_{\text{B}}} \xi_{\text{r,A}}^{\text{s}} - \frac{f_{\text{B}}}{f_{\text{A}} + f_{\text{B}}} \xi_{\text{r,B}}^{\text{s}} \\
 & + \frac{f_{\text{A}}}{f_{\text{A}} - f_{\text{B}}} \epsilon_{\text{r,A}}^{\text{s}} - \frac{f_{\text{B}}}{f_{\text{A}} - f_{\text{B}}} \epsilon_{\text{r,B}}^{\text{s}} \\
 & - \frac{f_{\text{A}}}{f_{\text{A}} + f_{\text{B}}} e_{\text{r,A}}^{\text{s}} - \frac{f_{\text{B}}}{f_{\text{A}} + f_{\text{B}}} e_{\text{r,B}}^{\text{s}} .
 \end{aligned} \tag{20.33}$$

The terms in (20.33) are the wide-lane ambiguity, the pseudorange and carrier-phase biases, the group-delay variations and phase-center variations, and the measurement errors due to noise and multipath. Therefore the Melbourne–Wübbena combination yields a biased estimate of the wide-lane ambiguity. It is affected by the combined noise of carrier-phase and pseudorange. The latter is the dominant source of noise although its contribution will be reduced through the narrow-lane combination. The Melbourne–Wübbena combination is commonly used for quality control of carrier-phase observations to detect cycle slips [20.17].

The narrow- and wide-lane combinations introduced here have been confined to using signals from only two frequencies. However, narrow-lane, intermediate-lane, or wide-lane combinations can of course also be formed by using observations on three or more frequencies simultaneously. A variety of different

combinations with different characteristics with respect to noise amplification and suppression of ionospheric delays has been introduced in the literature [20.3, 18, 19]. The references cited here may serve as a starting point for the reader to find more information about these combinations.

20.2.2 Ionosphere Combination

Since the ionospheric delay depends on the frequency of the observable, the first-order delay can be isolated with dual-frequency pseudorange or carrier-phase measurements. Based on (20.9) and (20.10), coefficients can be found that eliminate geometry ($h_1 = 0$) and preserve the ionosphere ($h_2 = 1$) using only pseudorange observations. Using these conditions and solving for the coefficients yields

$$\beta_A = -\beta_B = -\frac{f_B^2}{f_A^2 - f_B^2}. \quad (20.34)$$

Substituting these coefficients into (20.8) and including all terms of (20.2) yields

$$\begin{aligned} & -\frac{f_B^2}{f_A^2 - f_B^2} (p_{r,A}^s - p_{r,B}^s) \\ &= I_{r,A}^s(t) + \frac{f_B^2}{f_A^2 - f_B^2} c d_{r,AB}^s \\ & \quad - \frac{f_B^2}{f_A^2 - f_B^2} (\xi_{r,A}^s(t) - \xi_{r,B}^s(t)) \\ & \quad - \frac{f_B^2}{f_A^2 - f_B^2} (e_{r,A}^s(t) - e_{r,B}^s(t)), \end{aligned} \quad (20.35)$$

where the notation

$$d_{r,AB}^s = d_{r,B}^s - d_{r,A}^s$$

has been used for the combined satellite- and receiver-dependent differential code bias (DCB). It becomes obvious that the scaled difference of two pseudorange observables yields the first-order ionospheric delay on f_A biased by the DCB of the corresponding signals. It is also affected by the combined group-delay variations, multipath, and noise.

In the same way the coefficients for carrier-phase observations can be found

$$\alpha_A = -\alpha_B = \frac{f_B^2}{f_A^2 - f_B^2}. \quad (20.36)$$

Note the different sign of the coefficients for pseudorange and carrier-phase. These coefficients yield after

substitution into (20.8) using all terms of (20.4)

$$\begin{aligned} & \frac{f_B^2}{f_A^2 - f_B^2} (\varphi_{r,A}^s - \varphi_{r,B}^s) \\ &= I_{r,A}^s(t) + \frac{f_B^2}{f_A^2 - f_B^2} c (\delta_{r,A}^s - \delta_{r,B}^s) \\ & \quad + \frac{f_B^2}{f_A^2 - f_B^2} (\lambda_A N_{r,A}^s - \lambda_B N_{r,B}^s) \\ & \quad + \frac{f_B^2}{f_A^2 - f_B^2} (\zeta_{r,A}^s(t) - \zeta_{r,B}^s(t)) \\ & \quad + \frac{f_B^2}{f_A^2 - f_B^2} (\lambda_A - \lambda_B) \omega_r^s(t) \\ & \quad + \frac{f_B^2}{f_A^2 - f_B^2} (\epsilon_{r,A}^s(t) - \epsilon_{r,B}^s(t)). \end{aligned} \quad (20.37)$$

Similarly to the pseudorange, the scaled difference of the dual-frequency carrier-phase observations yields the ionospheric delay biased by the combined carrier-phase biases and ambiguities. In addition, the difference of the phase-center offset and variations, the phase wind-up and the carrier-phase noise and multipath are present in the observation equation.

The standard deviation of the measurement noise for the ionosphere combination σ_{IC} can be obtained from (20.18) assuming identical noise for both observations used in (20.35) or (20.37)

$$\sigma_{IC} = \sqrt{2} \frac{f_B^2}{f_A^2 - f_B^2} \sigma. \quad (20.38)$$

Figure 20.1 shows the ionosphere combination of pseudoranges and carrier-phases for two GPS satellites. The different noise level of both observations becomes immediately obvious from the plots. It is also interesting to note that neither the pseudorange nor the carrier-phase combination represent the true slant ionospheric delay. The pseudorange combination is offset by the biases of the observations. This causes the combination of one of the satellite to be negative, physically impossible for the true ionospheric delay. The same reasoning is also true for the carrier-phase combination, where, in addition to the biases, the combined ambiguities can cause arbitrary offsets. It can also be seen in the plot how the slant delays change with satellite elevation and with the diurnal change in local ionospheric activity at the receiver site. The ionosphere is rather calm at the beginning and becomes more and more active toward the end.

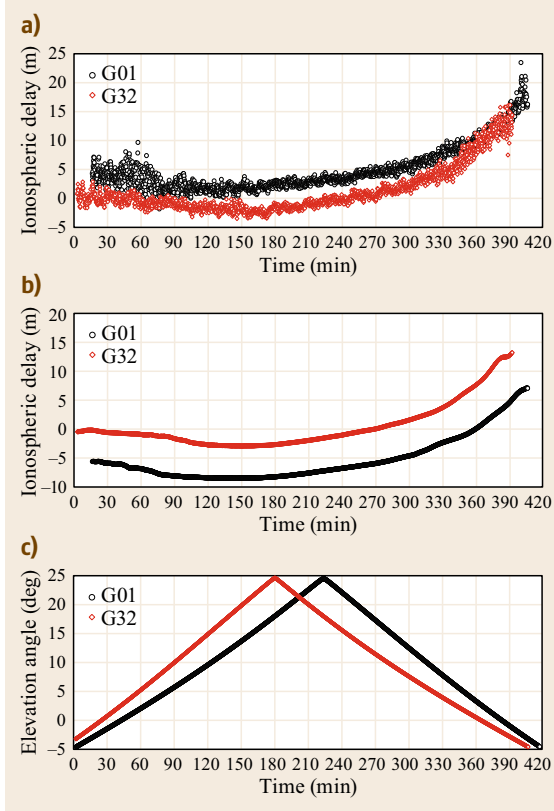


Fig. 20.1a–c Plot of slant ionospheric delay based on pseudorange (a) and carrier-phase (b) from L1 C/A and L2 P(Y) observations for two GPS satellites and corresponding satellite elevation angle (c)

20.2.3 Ionosphere-Free Combination

This section introduces signal combinations which are frequently used to eliminate the signal delay originating from the ionosphere. If only single-frequency observations are available, the GRAPHIC combination can be formed. The acronym GRAPHIC stands for group and phase ionospheric calibration [20.20]. Since it only requires pseudorange and carrier-phase observations on one frequency, it is even suitable for use with inexpensive single-frequency receivers. The GRAPHIC combination exploits the opposite sign of the ionospheric error in the pseudorange and carrier-phase and is formed from

$$\begin{aligned}
 o_{\text{GPH}(p_j, q_j)} &= \frac{1}{2} (p_{r,j}^s + \varphi_{r,j}^s) \\
 &= P_{r,\text{IF}}^s + \frac{1}{2} c (d_{r,j}^s + \delta_{r,j}^s) + \frac{1}{2} (\xi_{r,j}^s(t) + \zeta_{r,j}^s(t)) \\
 &\quad + \frac{1}{2} \lambda_j (\omega_r^s(t) + N_{r,j}^s) + \frac{1}{2} (e_{r,j}^s(t) + \epsilon_{r,j}^s(t)).
 \end{aligned} \tag{20.39}$$

The factors of the combination satisfy the condition for preservation of geometry $h_1 = 1$ and elimination of ionosphere $h_2 = 0$. In addition to the elimination of the error due to the ionosphere, all pseudorange errors and biases are reduced by a factor of 0.5. However, the new observable is now also affected by all nuisance parameters of the carrier-phase observable weighted with a factor of 0.5. Most significant is the effect of the carrier-phase ambiguity N in (20.39), which requires to introduce it as additional estimation parameter in positioning applications. As a result, the GRAPHIC combination must be treated in the processing similar to a carrier-phase observation; however, its measurement noise is dominated by the larger contribution of the pseudorange.

With observations on two or more different frequencies, more sophisticated ionosphere-free signal combinations can be formed. They exploit the fact that the ionosphere is a dispersive medium and thus observations on different frequencies are affected differently by the corresponding delay. With at least two observations on different frequencies, it is thus possible to determine the delay and eliminate it from the observation equations, at least to first order.

To begin with, the well-established dual-frequency ionosphere-free combination is introduced. It is formed using two pseudorange or two carrier-phase observations on different frequencies. Contrary to the ionosphere combination in the previous section where the ionospheric delay was isolated, this time a combination is desired which preserves geometry ($h_1 = 1$) and eliminates the ionosphere ($h_2 = 0$). The coefficients β_A and β_B for a dual-frequency pseudorange combination can then be obtained from (20.9) and (20.10)

$$\beta_A = 1 - \beta_B = \frac{f_A^2}{f_A^2 - f_B^2}, \tag{20.40}$$

where pseudoranges on the frequencies A and B are used in the combination. This leads to the ionosphere-free pseudorange combination equation

$$p_{r,\text{IF}}^s = \frac{f_A^2}{f_A^2 - f_B^2} p_{r,A}^s - \frac{f_B^2}{f_A^2 - f_B^2} p_{r,B}^s. \tag{20.41}$$

With the same coefficients, an ionosphere-free carrier-phase combination can also be formed as follows

$$\varphi_{r,\text{IF}}^s = \frac{f_A^2}{f_A^2 - f_B^2} \varphi_{r,A}^s - \frac{f_B^2}{f_A^2 - f_B^2} \varphi_{r,B}^s. \tag{20.42}$$

The advantage of eliminating the error of the first-order ionospheric delay comes at a price, however: the standard deviation of the combined observable will be significantly increased compared to the original uncombined observations. If identical noise σ is assumed for

both observations in the combination, (20.18) yields

$$\sigma_{\text{IF}} = \sqrt{\frac{f_A^2 + f_B^2}{f_A^2 - f_B^2}} \sigma, \quad (20.43)$$

where σ_{IF} is the noise of the ionosphere-free combination. The noise increases by a factor of approximately 3.0 for GPS L1/L2 signals, and by about 2.6 for signals on the GPS L1/L5 or Galileo E1/E5a bands [20.1]. It becomes obvious that the noise is lower when the signal bands have a larger separation in frequency. Combining observations on frequency bands closely together would lead to an even higher increase in noise, for example, by a factor of ≈ 16.6 for GPS L2/L5 or even ≈ 27.5 for Galileo E5a/E5b. Of course, the noise amplification in (20.43) holds for both pseudorange and carrier-phase observations.

For the ionosphere-free carrier-phase combination the resulting wavelength is of interest. The ionosphere-free condition $h_2 = 0$ yields from (20.10) that

$$\frac{\alpha_A}{f_A^2} + \frac{\alpha_B}{f_B^2} = 0. \quad (20.44)$$

Substituting the integer coefficients from (20.12) and the expression for the base frequency from (20.20) yields the following condition

$$\frac{i_A}{i_B} = -\frac{f_A}{f_B} = -\frac{k_A}{k_B}. \quad (20.45)$$

Noting from the geometry-preserving conditions that $\sum \alpha_j = 1$, the corresponding wavelength of the ionosphere-free combination can be found from (20.15)

$$\lambda_{\text{IF}} = \frac{\lambda_A \lambda_B}{i_A \lambda_B + i_B \lambda_A}. \quad (20.46)$$

It becomes obvious from (20.45) that multiple choices are possible for the integer coefficients. One option is to set the integers equal but with opposite sign to the base frequency multiplication factors k_j , divided by their lowest common denominator. This choice also yields the largest wavelength. Substituting numerical values for the GPS signals on L1 and L2 into (20.46) yields a wavelength of only approximately 6 mm, hampering ambiguity resolution. Non-integer choices for i_j , which satisfy (20.46), destroy the integerness of the resulting combined ambiguity.

For modeling the carrier-phase wind-up correction it is helpful to define the wavelength $\lambda_{\text{IF, PWU}}$ as follows

$$\lambda_{\text{IF, PWU}} = \lambda_A \frac{f_A^2}{f_A^2 - f_B^2} - \lambda_B \frac{f_B^2}{f_A^2 - f_B^2} = \frac{c}{f_A + f_B}, \quad (20.47)$$

which is identical to the narrow-lane wavelength (compare (20.26) and Table 20.1). It is worthwhile to emphasize that the dual-frequency ionosphere-free combinations for pseudorange and carrier-phase are only constructed from observations of the same type. Thus, in contrast to the GRAPHIC combination, they maintain their basic characteristic properties and can be processed similarly to the original uncombined observations except for the increased noise.

With dual-frequency measurements, only one ionosphere-free combination can be formed. With signals on more than two frequencies available, several possibilities to eliminate the ionosphere are possible and preserve the integer nature of the combined ambiguity [20.21, 22]. It is also possible to completely eliminate second-order ionospheric effects from the observations. Under normal ionospheric conditions, this error is on the order of a few centimeters in zenith direction for single-frequency observations and is already reduced by a factor of $\approx 30\%$ in the dual-frequency combination. With triple-frequency observations, the ionosphere can be completely eliminated up to second order, leaving only third-order residual errors, which then typically amount to less than a millimeter in this combination [20.23]. However, the noise amplification of these combinations is large when only L-band signals are used, which may limit their practical use [20.24].

It should also be noted that the use of multi-frequency signal combinations has been suggested, which do not completely eliminate the first-order ionosphere, but reduce it to a level where the effect becomes tolerable or negligible for ambiguity resolution. They are often referred to as ionosphere-reduced signal combinations [20.3, 25, 26]. The great variety of these new combinations prohibits their detailed discussion in this section. The reader may refer to specialized literature for further information on these combinations.

20.2.4 Multipath Combination

Multipath effects are mostly considered a nuisance in GNSS measurements and it is important to have an understanding of the magnitude of this error on the observations (Chap. 15). Multipath combinations can be used for this purpose.

The most frequently used multipath combination to assess pseudorange multipath on a single-frequency observation consists of one pseudorange observations on frequency A and two carrier-phase observations on frequencies A and B [20.27]

$$o_{\text{MP}(P_A)} = p_{r,A}^s - \varphi_{r,A}^s - 2k (\varphi_{r,A}^s - \varphi_{r,B}^s), \quad (20.48)$$

with k defined as

$$k = \frac{f_B^2}{f_A^2 - f_B^2}. \quad (20.49)$$

It becomes obvious from (20.9) and (20.10) that this multipath combination is geometry-free and ionosphere-free. Note the similarity of the last term in (20.48) to (20.37): essentially, this multipath combination is a code-carrier-difference minus twice the ionospheric delay. The remaining terms in the resulting combined observation $o_{MP(p_A)}$ are group-delay variations, signal biases, receiver noise, and multipath of the pseudorange observables, as well as a combination of phase-center variation and offset, signal biases, phase wind-up, ambiguities, noise, and multipath of the carrier-phase observable. The effect of phase wind-up, however, is reduced by a factor of 0.10–0.15 and amounts to only a few centimeters.

Assuming, firstly, that multipath and noise of the carrier-phase observations are negligible compared to the corresponding terms of the pseudorange and, secondly, that all other terms are constant, the evaluation of the multipath combination over time yields the temporal variation of this nuisance parameter biased by a combination of the remaining terms

$$\begin{aligned} o_{MP(p_A)} = & e_{r,A}^s(t) - ((1+2k)\lambda_A + 2k\lambda_B)\omega_r^s(t) \\ & + \xi_{r,A}^s(t) - (1+2k)\zeta_{r,A}^s(t) + 2k\zeta_{r,B}^s(t) + \Gamma, \end{aligned} \quad (20.50)$$

with the constant term Γ defined as

$$\begin{aligned} \Gamma = & cd_{r,A}^s - (1+2k)c\delta_{r,A}^s + 2kc\delta_{r,B}^s \\ & - (1+2k)\lambda_A N_{r,A}^s + 2k\lambda_B N_{r,B}^s. \end{aligned} \quad (20.51)$$

According to the previous assumption, the noise of this multipath combination is dominated by the noise of the pseudorange measurement. However, for special low-noise pseudorange signals, like Galileo alternative BOC (AltBOC), both code and phase noise may be relevant.

An example of this multipath combination is shown in Fig. 20.2. The figure depicts the multipath errors of C/A-code pseudorange measurements from a GPS satellite over time tracked by a geodetic grade receiver. After the satellite elevation angle drops below 12° at 900 s the multipath errors become clearly visible. The effect can also be observed in the periodic oscillations of the measured C/N_0 , typical for the alternating destructive and constructive interference of multipath signals [20.28]. At elevation angles less than 6° , the multipath becomes so severe that the receiver loses lock repeatedly and the ambiguities in (20.51) change.

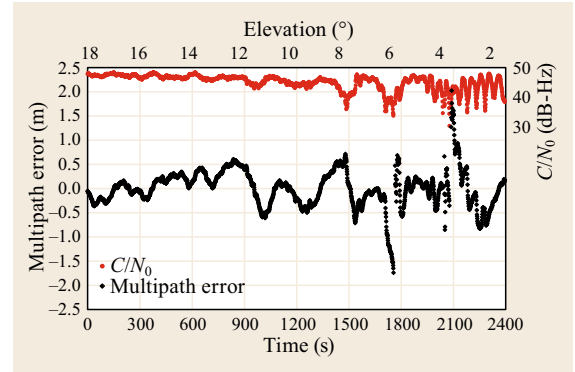


Fig. 20.2 Plot of multipath combination (black) and measured carrier-to-noise-power-density ratio C/N_0 (red) over time and satellite elevation angle for L1 C/A code measurements of a GPS satellite. Note that the change of the elevation angle is approximately linear over the plot interval

The previous derivations have shown how multipath on a single pseudorange measurement can be assessed with the help of two carrier-phase observations on different frequencies. With a combination of observations from three frequencies, isolation of the multipath errors is possible to some extent for pseudorange and carrier-phase observations separately [20.29]. Consider the following triple-frequency combinations for pseudorange $o_{MP(p_A, p_B, p_C)}$ and carrier-phase $o_{MP(\phi_A, \phi_B, \phi_C)}$

$$o_{MP(p_A, p_B, p_C)} = \beta_A p_{r,A}^s + \beta_B p_{r,B}^s + \beta_C p_{r,C}^s \quad (20.52)$$

and

$$o_{MP(\phi_A, \phi_B, \phi_C)} = \alpha_A \phi_{r,A}^s + \alpha_B \phi_{r,B}^s + \alpha_C \phi_{r,C}^s, \quad (20.53)$$

where the linear coefficients are defined as

$$\begin{aligned} \alpha_A = \beta_A &= \frac{(\lambda_C^2 - \lambda_B^2)}{\Lambda}, \\ \alpha_B = \beta_B &= \frac{(\lambda_A^2 - \lambda_C^2)}{\Lambda}, \\ \alpha_C = \beta_C &= \frac{(\lambda_B^2 - \lambda_A^2)}{\Lambda}, \\ \Lambda^2 &= (\lambda_C^2 - \lambda_B^2)^2 + (\lambda_A^2 - \lambda_C^2)^2 + (\lambda_B^2 - \lambda_A^2)^2. \end{aligned} \quad (20.54)$$

Evaluating the geometry scaling factor (20.9) and the ionospheric scaling factor (20.10) with the coefficients (20.54) reveals that this combination is both geometry- and ionosphere-free. Furthermore, evaluation of (20.18) yields a noise-amplification factor of one

assuming that all three observables have identical noise levels [20.30]. Substituting (20.2) into (20.52) yields for pseudorange

$$\begin{aligned} \text{OMP}(p_A, p_B, p_C) &= \beta_A e_{r,A}^s(t) + \beta_B e_{r,B}^s(t) + \beta_C e_{r,C}^s(t) \\ &+ \beta_A \xi_{r,A}^s(t) + \beta_B \xi_{r,B}^s(t) + \beta_C \xi_{r,C}^s(t) \\ &+ c(\beta_A d_{r,A}^s + \beta_B d_{r,B}^s + \beta_C d_{r,C}^s). \end{aligned} \quad (20.55)$$

The corresponding factors for open-service signals of the new and modernized GNSS can be found in Table 20.2. It becomes obvious that the terms associated with the first, highest frequency are attenuated compared to the other two frequencies, which have similar weights. For pseudorange, the combination depends on the combined multipath errors and receiver noise $e_{r,j}^s$, a combination of the individual group-delay variations $\xi_{r,j}^s$, and a combination of the pseudorange delays $d_{r,j}^s$. Although the latter are generally assumed constant, investigations with triple-frequency multipath combinations suggest that frequency-dependent line-bias variations exist for some satellites [20.31].

Figure 20.3 shows the triple-frequency pseudorange combination with GPS L1 C/A, L2C, and L5 observations for the same satellite and time period as in Fig. 20.2. Again, the multipath effects are clearly visible at low satellite elevation angles.

Similar to pseudorange, the triple-frequency carrier-phase multipath combination is found from substituting (20.4) into (20.53)

$$\begin{aligned} \text{OMP}(\phi_A, \phi_B, \phi_C) &= \alpha_A \epsilon_{r,A}^s(t) + \alpha_B \epsilon_{r,B}^s(t) + \alpha_C \epsilon_{r,C}^s(t) \\ &+ \alpha_A \zeta_{r,A}^s(t) + \alpha_B \zeta_{r,B}^s(t) + \alpha_C \zeta_{r,C}^s(t) \\ &+ c(\alpha_A \delta_{r,A}^s + \alpha_B \delta_{r,B}^s + \alpha_C \delta_{r,C}^s) \\ &+ \omega_r^s(\alpha_A \lambda_A + \alpha_B \lambda_B + \alpha_C \lambda_C) \\ &+ \alpha_A \lambda_A N_{r,A}^s + \alpha_B \lambda_B N_{r,B}^s + \alpha_C \lambda_C N_{r,C}^s. \end{aligned} \quad (20.56)$$

Table 20.2 Coefficients for selected triple-frequency multipath combinations of different GNSSs. The last column lists the effective wavelength of the phase wind-up term

GNSS	α_j, β_j			$\sum \alpha_j \lambda_j$ (mm)
GPS, QZSS L1,L2,L5	0.142	-0.767	0.626	-0.99
Galileo E1,E5b,E5a	0.085	-0.746	0.661	-0.63
GLONASS L1,L2,L3	0.121	-0.760	0.639	-0.82
BeiDou B1,B3,B2	0.183	-0.781	0.597	-0.94

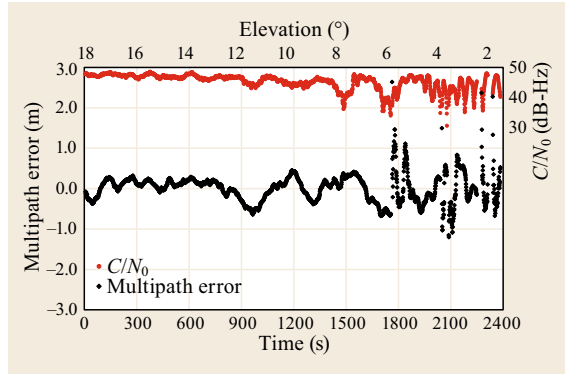


Fig. 20.3 Plot of triple-frequency multipath combination (black) for L1 C/A, L2C, and L5 pseudorange observations and measured L1 C/A carrier-to-noise-power-density ratio C/N_0 (red) over time and satellite elevation angle of a GPS satellite. Note that the change of the elevation angle is approximately linear over the plot interval

Besides the combined effect of multipath and noise $\epsilon_{r,j}^s$, the carrier-phase combination depends on the combinations of phase-center variations $\zeta_{r,j}^s$, the combined phase biases $\delta_{r,j}^s$, a linear combination of the ambiguities $N_{r,j}^s$, as well as the phase wind-up ω_r^s . The last column in Table 20.2 lists the combined multiplication factors for this term. It is on the order of a millimeter and thus virtually eliminated from the carrier-phase combination.

Figure 20.4 shows the triple-frequency carrier-phase combination with GPS observations of the L1 C/A, L2C, and L5 signals for the same satellite and time period as in the previous two multipath plots. The effect

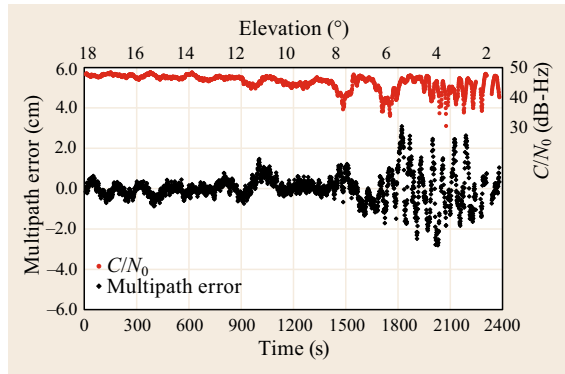


Fig. 20.4 Plot of multipath combination (black) for a triple-frequency carrier-phase combination of L1 C/A, L2C, and L5 observations and measured L1 C/A carrier-to-noise-power-density ratio C/N_0 (red) over time and satellite elevation angle of a GPS satellite. Note that the change of the elevation angle is approximately linear over the plot interval

of the combined multipath becomes visible in the high-frequency oscillations at low elevation angles. Note that the ordinate axis has a different scale compared to

Fig. 20.3, since the carrier-phase multipath is two orders of magnitude smaller. For a detailed discussion of multipath effects, the reader may refer to Chap. 15.

20.3 Combinations of Multisatellite and Multireceiver Observations

In the previous sections observations combinations have only been formed from measurements of a single satellite tracked by a single receiver. Of course, combinations can also be formed between two satellites or between two receivers. The latter is a long-established technique for differential positioning and allows for convenient elimination of nuisance parameters from the observables. Differences of observations of two satellites at one receiver or of two receivers for the same satellite are called single differences. Taking the difference between two single differences leads to a double-difference observable. Finally, the time difference between two double differences at different epochs is referred to as triple difference. At the end of the section, the special case of a zero-baseline configuration will be introduced, where two receivers are connected through a power-splitter to the same antenna. Single-, double-, and triple differences with this setup are very useful for the characterization of GNSS signals or receiver measurements.

A short note on notation shall be added at this point: so far all time-dependent terms have succeeded by (t) . For the remainder of the chapter, this indicator for time dependency is dropped to achieve a more concise notation.

20.3.1 Between-Receiver Single Difference

Assume that time-synchronized measurements of the same satellite k from two receivers 1 and 2 as depicted in Fig. 20.5 are available. The between-receiver single differences (BRSD) p_{12}^k of the pseudorange observations are then computed from

$$p_{12,\boxtimes}^k = p_{2,\boxtimes}^k - p_{1,\boxtimes}^k. \quad (20.57)$$

Note the symbol \boxtimes in place of the signal index j . This notation has been chosen here to indicate that the single differences are typically formed between identical signals, but not necessarily basic single observations. One could, for example, also use an ionosphere-free observable or any other signal combination in the single difference depending on the desired application. For the remainder of this section, only single, uncombined measurements will be considered and the placeholder \boxtimes will therefore be dropped for brevity. Substitut-

ing (20.2) into (20.57) and considering all the terms yields for the single-difference pseudorange

$$p_{12}^k = \rho_{12}^k + c \left(dt_{12} + \delta t_{\text{stc},12}^{\text{rel},k} \right) + cd_{12}^k + \xi_{12}^k + T_{12}^k + I_{12}^k + e_{12}^k. \quad (20.58)$$

The brief notation $\rho_{12}^k = \rho_2^k - \rho_1^k$ has been used for the differential geometric range and equivalently also for all other terms. A different notation can also often be found in literature, which denotes a between-receiver single difference as Δp_{12}^k and is equivalent to (20.57). However, in this case an extra Δ would be included in the observation equation for each term, which can lead to a rather cumbersome notation.

The term for the satellite clock offset, which is of course identical for two time-synchronized observations at different receivers irrespective of the antenna distance, has dropped out of (20.58). The same also holds true for the relativistic correction for the satellite clock due to noncircular orbits (19.15) and the J_2 -correction (19.18). The terms for the differential relativistic space-time curvature correction $\delta t_{\text{stc},12}^{\text{rel},k}$, tropospheric delay T_{12}^k , ionospheric delay I_{12}^k , and group delay variation ξ_{12}^k are still present in the equations, but strongly correlate with the distance of the two antennas. For antenna distances of only up to a few 100 m, the signal transmit paths in the atmosphere are virtually the same, which also leads to identical delays of the signals and these terms drop out of (20.58). For larger antenna separations, the electron content in the ionosphere and

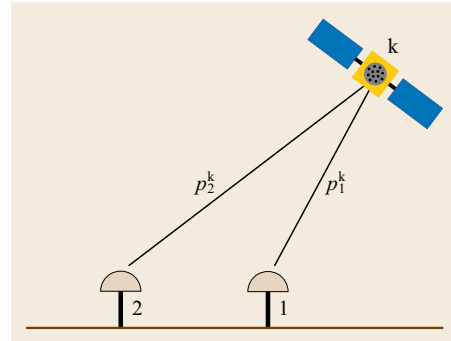


Fig. 20.5 Single difference of observations from satellite k between receivers 1 and 2

the atmospheric parameters in the troposphere may differ and become significant. A similar reasoning also holds true for the differential group delay variation ξ_{12}^k , since for small antenna baseline lengths, the azimuth and off-boresight angle of the antenna-satellite line-of-sight vectors are practically identical, thus the signal is affected by the same group delays. For larger baselines the group delays may differ. The relativistic correction due to space-time curvature (19.14) also depends on the line-of-sight vector, and thus the term $\delta_{\text{stc},12}^{\text{rel},k}$ decorrelates for large baselines.

The differential biases d_{12}^k have been retained in (20.58). The reason for this is that only if receivers with identical correlators are used in the single difference, the combined satellite and receiver bias can be split up into the individual terms as in (20.3). In this case, the satellite-dependent part drops out of the single difference, whereas the differential receiver-dependent part is retained. Finally, the terms for differential receiver clock offset dt_{12} and noise and multipath errors e_{12}^k remain in the equations irrespective of the antenna separation.

In a similar manner, a single-difference carrier-phase between two receivers φ_{12}^k can be formed

$$\begin{aligned} \varphi_{12}^k = & \rho_{12}^k + c \left(dt_{12} + \delta_{\text{stc},12}^{\text{rel},k} \right) + c \delta_{12}^k + \zeta_{12}^k \\ & + T_{12}^k - I_{12}^k + \lambda \left(\omega_{12}^k + N_{12}^k \right) + \epsilon_{12}^k. \end{aligned} \quad (20.59)$$

In (20.59) the differential terms for the range ρ_{12}^k , the receiver clock dt_{12} , the relativistic clock correction $\delta_{\text{stc}}^{\text{rel}}$, the differential tropospheric delay T_{12}^k , the ionospheric delay I_{12}^k , the phase-center variation ζ_{12}^k , the carrier-phase biases δ_{12}^k , and the noise and multipath ϵ_{12}^k are identical or similar to the single difference pseudorange equation and their discussion, therefore,

not repeated here. Two additional terms have appeared in (20.59): the differential phase wind-up ω_{12}^k and the single-difference ambiguity N_{12}^k .

The phase wind-up term depends on the relative orientation of the transmitting and receiving antennas. Assuming that the orientation of the receiving antennas does not change with respect to each other, the residual phase wind-up effect in differential measurements over short baselines on the order of a few 100 km or less will be on the order of a few millimeters and may thus be neglected. The residual effects increase with increasing baseline length and can theoretically amount to half a wavelength for receiving antennas on opposite sides of the Earth [20.32]. The receiving antennas may rotate with respect to each other, for example, in case of one antenna being setup as a reference station and the other antenna being mounted on a moving vehicle. In this case, the change in the differential phase wind-up between the two antennas must also be accounted for.

Figure 20.6 shows the between-receiver single differences for pseudorange observations Fig. 20.6a and carrier-phase observations Fig. 20.6b of three GPS satellites over an interval of 4 h. The two antennas are mounted at a distance of 4.8 m, thus differential troposphere, ionosphere, and group-delay variations are eliminated. The remaining terms are the differential range, the differential receiver clock offset and bias, the differential relativistic effects due to space-time curvature, and differential receiver noise and multipath errors. The carrier-phase single differences are also arbitrarily offset by the differential ambiguity. In this case, however, both receivers keep their carrier-phase observations closely aligned to the pseudorange. Therefore the ordinates of both plots still have the same scale. It becomes immediately obvious from the

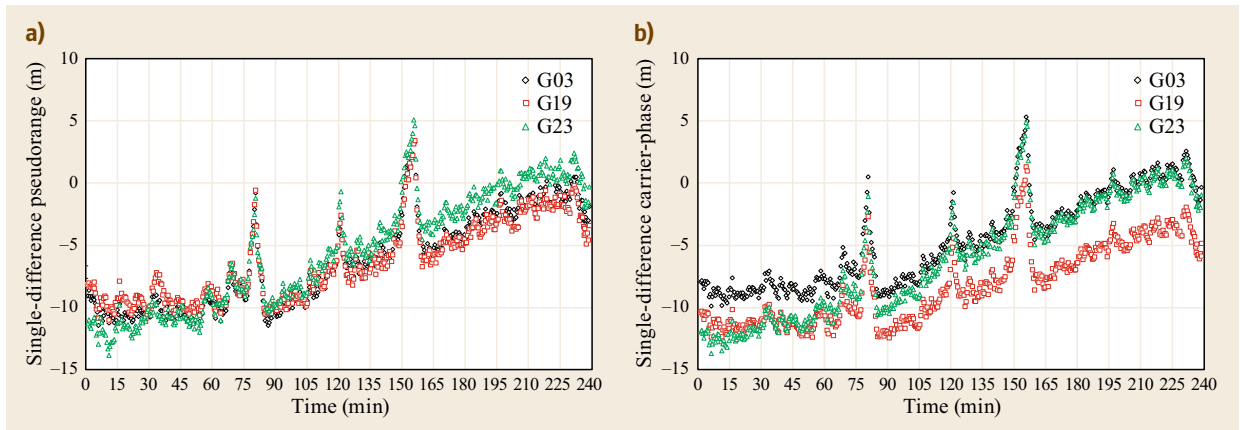


Fig. 20.6a,b Between-receiver single differences of pseudorange (a) and carrier-phase (b) observations for three GPS satellites and two antennas with a baseline length of 4.8 m over a time interval of 4 h

plots that both pseudorange and carrier-phase single differences exhibit identical short-term temporal variations, which are especially pronounced at 75, 120, and 150 min. These variations are due to the differential receiver clock offsets, since the receivers have been operated using their independent internal oscillators. In addition, the change of differential range over time also contributes to an identical systematic variation of both observables over time. Since it only depends on the satellite motion, this change is slow in time for a static baseline. At maximum it can only vary over twice the baseline length, in this case from -4.8 to $+4.8$ m.

20.3.2 Between-Satellite Single Difference

The single-difference equations (20.58) and (20.59) have been formed between two receivers using observations of the same satellite. Of course, single differences

$$\begin{aligned} p_1^{kl} &= p_1^l - p_1^k, \\ \varphi_1^{kl} &= \varphi_1^l - \varphi_1^k, \end{aligned} \quad (20.60)$$

can also be formed using two observations of different satellites k and l at the same receiver 1 as shown in Fig. 20.7. In case the more elaborate notation is used, the between-satellite single differences (BSSD) are denoted ∇p_1^{kl} and $\nabla \varphi_1^{kl}$. Note that ∇ is used for BSSD instead of Δ for BRSD.

The observations equations for satellite-satellite single differences of pseudoranges p_1^{kl} and carrier-phases φ_1^{kl} are as follows

$$\begin{aligned} p_1^{kl} &= \rho_1^{kl} + c(dt^{kl} + \delta t^{\text{rel},kl}) + cd_1^{kl} + \xi_1^{kl} \\ &\quad + T_1^{kl} + I_1^{kl} + e_1^{kl}, \end{aligned} \quad (20.61)$$

$$\begin{aligned} \varphi_1^{kl} &= \rho_1^{kl} + c(dt^{kl} + \delta t^{\text{rel},kl}) + cd_1^{kl} + \zeta_1^{kl} \\ &\quad + T_1^{kl} - I_1^{kl} + \epsilon_1^{kl} \\ &\quad + \lambda^l (\omega_1^l + N_1^l) - \lambda^k (\omega_1^k + N_1^k). \end{aligned} \quad (20.62)$$

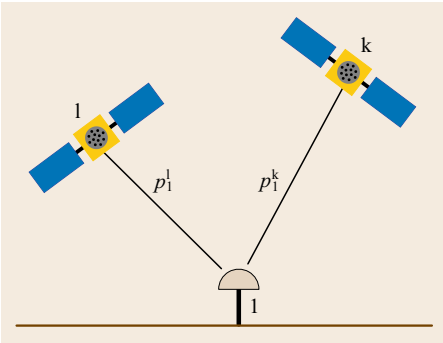


Fig. 20.7 Single difference of observations from receiver 1 between satellites k and l

The offset and the corresponding relativistic effect of the receiver clock drop out the equations for the between-satellite difference. All other terms are retained. In the carrier-phase equation (20.62), the expressions for the differences of phase wind-up and ambiguity are more complicated, since they may involve frequencies with different wavelengths λ^k and λ^l . When the single difference is formed between two satellites of different constellations, the ISB of these two constellations will manifest itself in the combined observable, which is the combination of a system-time offset and measurement biases. In (20.1) and (20.62), the contribution of the system-time offset is contained in the differential satellite clock offset dt^{kl} . The contribution of the biases in the ISB is contained in the differential pseudorange bias term d_1^{kl} and the differential carrier-phase bias term δ_1^{kl} .

Figure 20.8 depicts between-satellite single differences for pseudorange (Fig. 20.8a) and carrier-phase (Fig. 20.8b) for the same time interval as in Fig. 20.6 using data from one receiver only for the same three GPS satellites. The observations have been corrected for the differential geometric range and satellite clock offset, which can amount to several kilometers. As expected, the variations due to the change in the receiver clock have now vanished and the different levels of noise and multipath errors of the pseudorange and carrier-phase observations become distinguishable. Differential tropospheric and ionospheric delays, and differential group-delay variations and phase-center variations contribute to the temporal changes of the satellite-satellite single differences. Note the differences in the temporal variation between pseudorange and carrier-phase due to the different sign of the ionospheric delay. This effect is also referred to as code-carrier divergence.

As for most of the other combinations, the elimination of parameters from the measurement equations comes at the price of increased noise. The noise of the between-receiver single-difference σ_{12} or between-satellite single-difference σ^{kl} is

$$\sigma_{12} = \sigma^{kl} = \sqrt{2}\sigma, \quad (20.63)$$

assuming identical standard deviation σ for the uncombined observables.

20.3.3 Double Difference

A double-difference observation can be formed if observations from a pair of receivers 1 and 2 and a pair of satellites k and l are available as depicted in Fig. 20.9. For pseudorange measurements, it can be formed from two receiver-receiver single differences as follows

$$p_{12}^{kl} = p_{12}^l - p_{12}^k. \quad (20.64)$$

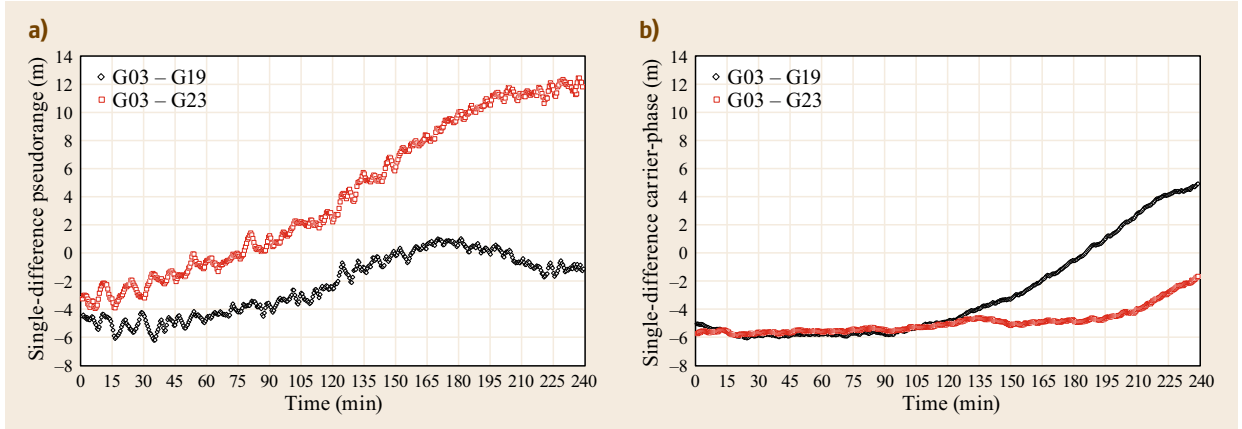


Fig. 20.8a,b Between-satellite single differences of pseudorange (a) and carrier-phase (b) observations for three GPS satellites over a time interval of 4 h. The observations have been corrected for the differential geometric range and the satellite clock offsets

Of course, the same result can also be obtained through the combination of two single differences between satellites $p_{12}^{kl} = p_2^{kl} - p_1^{kl}$. In the alternative notation, the double difference would be written as $\Delta \nabla p_{12}^{kl}$. Substituting (20.58) into (20.64) yields

$$p_{12}^{kl} = \rho_{12}^{kl} + c d_{12}^{kl} + \xi_{12}^{kl} + T_{12}^{kl} + I_{12}^{kl} + e_{12}^{kl}. \quad (20.65)$$

Note that the receiver and satellite clock offsets as well as the relativistic corrections have been eliminated from the double difference, but the double difference of the geometric range are still retained. The terms for the troposphere, the ionosphere, and the code-phase pattern variation are also still present in this most general form of the equation, but are canceled out for small antenna separations. Receiver noise and multipath e_{12}^{kl} are still be retained.

The bias double differences d_{12}^{kl} are present unless receivers with identical front-end characteristics and correlator settings are used. If satellites of different constellations are used in the double difference, parts of

the corresponding ISB are still retained in the combined observable. The system-time offset between the two constellations is of course identical for both receivers and is completely eliminated like the satellite clocks. However, the two receivers involved may exhibit different biases for different constellations. As a result, the double-difference biases of satellites from the same constellation may vanish, but can still be present when the two satellites are from different constellations.

The equivalent double-difference equation for carrier-phase observation φ_{12}^{kl} is given by

$$\begin{aligned} \varphi_{12}^{kl} = & \rho_{12}^{kl} + c \delta_{12}^{kl} + \xi_{12}^{kl} + T_{12}^{kl} - I_{12}^{kl} \\ & + \lambda (\omega_{12}^{kl} + N_{12}^{kl}) + \epsilon_{12}^{kl}. \end{aligned} \quad (20.66)$$

The double differences of geometric range, tropospheric, and ionospheric delays are already known from (20.65). The term ξ_{12}^{kl} describes the double difference of the phase-center variation. The double differences of the carrier-phase biases δ_{12}^{kl} are retained in the equations and will only drop out if receivers with compatible front-end and correlator design are used. The combined errors due to receiver noise and multipath are summarized in the double-difference term ϵ_{12}^{kl} . Finally, the phase wind-up term ω_{12}^{kl} and the ambiguity term N_{12}^{kl} are multiplied with the wavelength λ . It is implicitly assumed in this derivation that all carrier-phase observations involved in the double difference have the same wavelength. The phase wind-up term is included in the double difference, since this effect becomes negligible only for short baselines of a few 100 km. It can theoretically amount to half a wavelength for receiving antennas on opposite sides of the Earth. Only the contribution of a relative rotation of the receiving antennas cancels out for all baseline lengths in case of double dif-

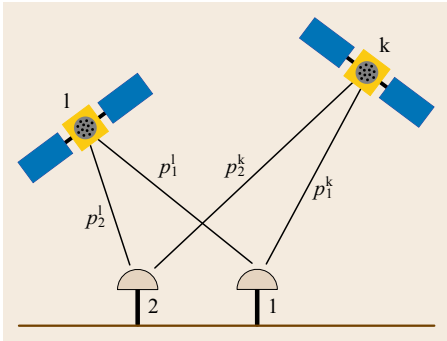


Fig. 20.9 Double difference of observations from receiver 1 and 2 and satellites k and l

ferences, since it is identical in both single difference observations involved in (20.66).

Figure 20.10 depicts double differences for pseudorange (Fig. 20.10a) and carrier-phase (Fig. 20.10b) for the same time interval using the same satellites and receivers as in the examples in the previous section. The short baseline length of 4.8 m eliminates all atmospheric delays and the effects due to group-delay variations and phase-center offset variation. Only the double difference of geometric range, signal biases, noise, and multipath remain. The change of ρ_{12}^{kl} amounts to a few meters and can clearly be recognized in the plots for both observables. The noise and multipath of the double-difference pseudoranges are also clearly visible. For the carrier-phases, these effects are much smaller and not distinguishable on this scale.

The noise of the double-difference equation is increased by a factor of $\sqrt{2}$ compared to the single-difference combination, since four observations with uncorrelated stochastic errors are involved. Using the expression for the single-difference standard deviation defined in (20.63) together with the error propagation law (20.18), one obtains

$$\sigma_{12}^{kl} = \sqrt{2}\sigma_{12} = \sqrt{2}\sigma^{kl} = 2\sigma, \quad (20.67)$$

for the measurement noise σ_{12}^{kl} the double differences.

20.3.4 Triple Difference

Forming single and double differences of observations has proven effective to eliminate some nuisance parameters, especially on short baselines where spatial correlation of signals delays can be exploited. Another method to eliminate errors is to take advantage of tem-

poral correlation and take the difference of observations at different epochs. The time difference of two double-difference observations is typically referred to as a triple difference [20.33]

$$\partial p_{12}^{kl} = p_{12}^{kl}(t_i) - p_{12}^{kl}(t_{i-1}). \quad (20.68)$$

The operator ∂ is used to denote the time difference of observations between the two epochs t_i and t_{i-1} . The time difference will eliminate all parameters in (20.66), which are constant over time. In the most general case, these are the double-difference terms for carrier-phase biases δ_{12}^{kl} and the carrier-phase double-difference ambiguity N_{12}^{kl} as long as no cycle slip occurs. The full triple-difference measurement equations including all terms are

$$\partial p_{12}^{kl} = \partial \rho_{12}^{kl} + \partial \xi_{12}^{kl} + \partial T_{12}^{kl} + \partial I_{12}^{kl} + \partial e_{12}^{kl}, \quad (20.69)$$

$$\partial \phi_{12}^{kl} = \partial \rho_{12}^{kl} + \partial \xi_{12}^{kl} + \partial T_{12}^{kl} - \partial I_{12}^{kl} + \partial \epsilon_{12}^{kl} + \partial \omega_{12}^{kl}. \quad (20.70)$$

Should the terms for the tropospheric and ionospheric delays, the phase-center variation, and the phase wind-up not be completely eliminated already in the double difference, their magnitude is further reduced in the triple difference, especially when observations of the time interval $t_i - t_{i-1}$ are small.

Figure 20.11 depicts triple differences for pseudorange (Fig. 20.11a) and carrier-phase (Fig. 20.11b). The triple differences have been formed from the double differences depicted in Fig. 20.10. Note the different scale of the ordinates for pseudorange and carrier-phase. Only the time differences of the geometric range and the receiver noise are still present in the triple-difference observations on the short baseline of 4.8 m.

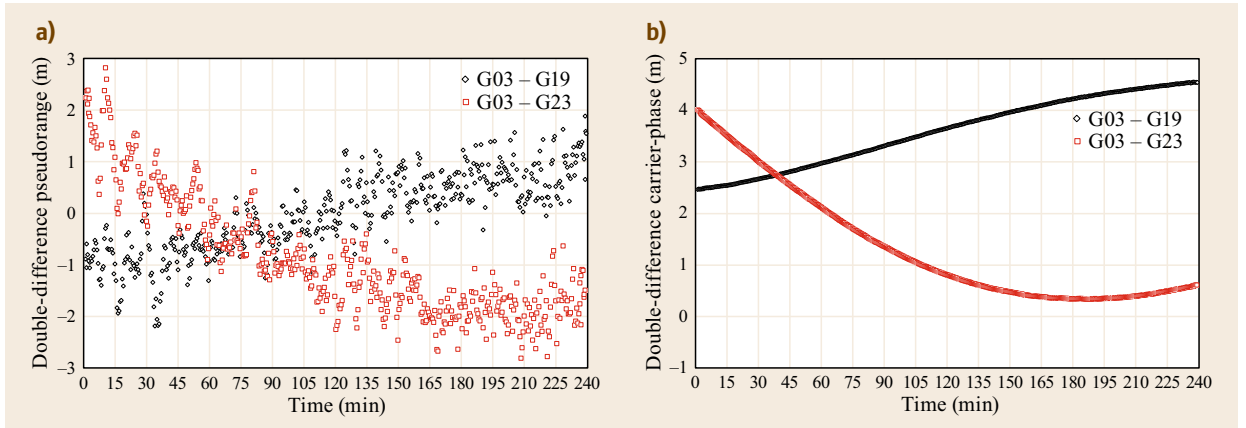


Fig. 20.10a,b Double differences of pseudorange (a) and carrier-phase (b) observations for three GPS satellites and two antennas with a baseline of 4.8 m over a time interval of 4 h

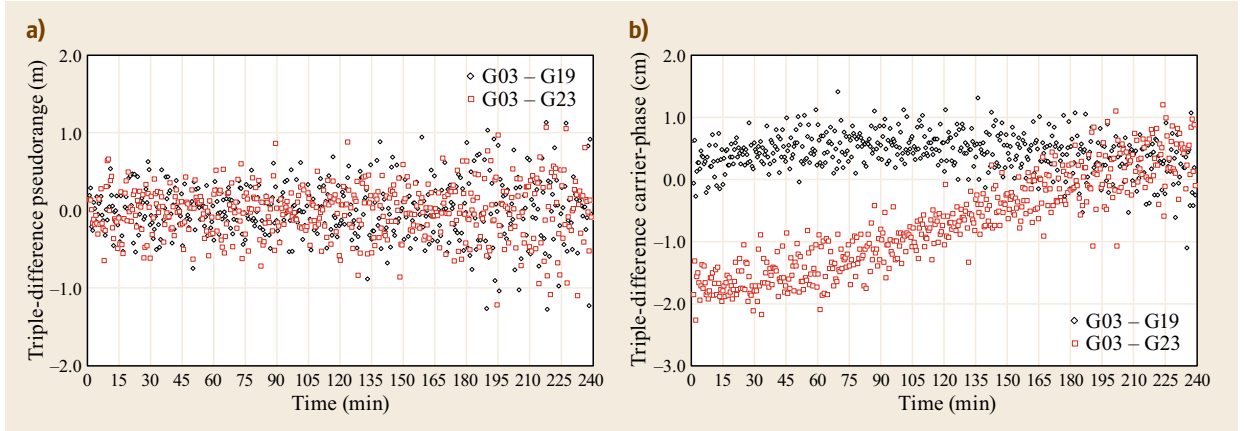


Fig. 20.11a,b Triple differences of pseudorange (a) and carrier-phase (b) observations for three GPS satellites and two antennas with a baseline length of 4.8 m over a time interval of 4 h. Note the different scales of the ordinates for pseudorange (m) and carrier-phase (cm)

Application of the error propagation law (20.18) and assuming equal noise for both double differences leads to the following expression for the noise of the triple-difference $\sigma_{\delta_{12}^{kl}}$

$$\sigma_{\delta_{12}^{kl}} = \sqrt{2}\sigma_{12}^{kl} = \sqrt{8}\sigma, \quad (20.71)$$

which uses (20.63) and (20.67) to relate to the noise of the individual observations.

20.3.5 Single and Double Difference on Zero-Baselines

A setup, where two or more receivers are connected via a power splitter to the same antenna, is commonly referred to as a zero-baseline configuration. It is depicted in Fig. 20.12. Since the signals are tracked by both receivers referred to the same antenna-phase center, the differential geometry, group-delay variations, phase-center offsets and variations, and the atmospheric delays are canceled completely in zero-baseline single differences. The only remaining terms are differential receiver clock offsets, biases, multipath, and receiver noise in case of pseudorange observations. For carrier-phase observations, the single-difference ambiguity is also retained. Dropping all terms in the single-difference observation equations (20.58) and (20.59), which cancel due to the identical signal transmission path, yields the following equations for the zero-baseline case

$$p_{12,ZB}^k = c(dt_{12} + d_{12}^k) + e_{12}^k, \quad (20.72)$$

$$\varphi_{12,ZB}^k = c(dt_{12} + \delta_{12}^k) + \lambda N_{12}^k + \epsilon_{12}^k. \quad (20.73)$$

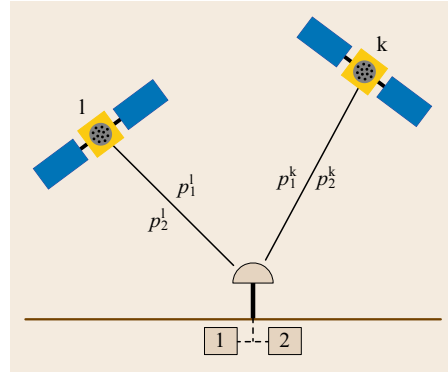


Fig. 20.12 Zero-baseline configuration of receiver 1 and 2 connected to the same antenna

Forming double differences further eliminates the receiver clock

$$p_{12,ZB}^{kl} = cd_{12}^{kl} + e_{12}^{kl}, \quad (20.74)$$

$$\varphi_{12,ZB}^{kl} = c\delta_{12}^{kl} + \lambda N_{12}^{kl} + \epsilon_{12}^{kl}. \quad (20.75)$$

Due to this convenient elimination of nuisance parameters, a zero-baseline configuration is a convenient setup for receiver characterization as well as signal studies.

The derived equations model the zero-baseline observations for the most general case using a pair of arbitrary receivers, which are not necessarily of the same type. Further terms may cancel when identical receiver models, or receivers with the same correlator and front-end configuration, are used in the zero-baseline configuration. In this case, also the pseudorange and carrier-phase bias terms will drop out, since the correlator will produce the same tracking error for the satellite's chip-shape distortions. For the same reason, the multipath errors, implicitly contained in the terms e_{12}^{kl} and ϵ_{12}^{kl} , will also cancel out in this case.

It should also be noted that the aforementioned full cancellation of the geometry term ρ_{12}^{kl} is only true for time-synchronized observations between the two receivers. The maximum range rate for a GPS satellite and a ground-based user at the equator is approximately 1000 m/s when the satellite is close to the horizon. If the observation epochs of two receivers in a zero-baseline setup for this satellites have a synchronization error of 1 ms, the satellite has moved approximately 1 m with respect to the user. This range change will then still be present in the differential measurement. For the analysis of carrier-phase observations it is desirable to confine the maximum range change to less than 1 mm, which requires the receiver clock synchronization errors to be less than 1 μ s. This can typically be achieved by activating the internal receiver clock synchronization to a GNSS system time or, in post-processing, by extrapolating the observations of the two receivers to a common epoch computed from pseudorange-based positioning solutions.

Figure 20.13 depicts single-differences of pseudorange observations for GPS, Galileo, and BeiDou satellites. The receivers are different models and have their internal clocks synchronized to GPS time within a few nanoseconds. This becomes obvious from the single differences of the three GPS satellites, which contain the differential receiver clock offset. It is on the order of a few meters and its variation over time is identical for all three satellites. In addition, differential receiver noise and multipath are present, but significant biases between the GPS satellites cannot be distinguished. Interestingly, the single differences of the BeiDou and the Galileo satellites between the two receivers exhibit a large offset compared to the GPS satellites. The reason for this offset is the difference of the ISB between these two receivers.

Even though the contribution of the system-time offset of the ISB is eliminated in the single difference, the differential biases between the different signals are still retained. In case of GPS and Galileo, the between-receiver ISB difference is ≈ 23 m. For GPS and BeiDou, it is on a similar order of magnitude, but not identical. Note that the temporal variation of the differential receiver clock is also visible in the Galileo and BeiDou single differences. The BeiDou satellite C12 exhibits increased differential noise and multipath errors at the beginning of the data arc, since it is tracked at very close to the horizon at an elevation angle of less than 10° for the first 15 min of the plotted data.

Figure 20.14 depicts the double differences of pseudorange and carrier-phase observations for the same time period, satellites, and receivers as in Fig. 20.13. A GPS satellite has been selected as reference satellite for all combinations. Inspection of the GPS double differences for pseudoranges in the top plot makes obvious that the differential clock offset and its variations are now removed and the double differences for GPS are centered at zero. As already expected, the mixed-constellation double differences with Galileo and BeiDou satellites are still offset due to the effect of the differential biases.

The carrier-phase double differences in the bottom plot of Fig. 20.14 are plotted in units of cycles. Note that the offset due to the integer ambiguity has been removed from the data. It becomes obvious that the GPS carrier-phase double differences are centered at zero, whereas the Galileo satellites exhibit an offset of -0.5 cycles and the BeiDou satellites have an offset of approximately 0.25 cycles. These offsets are the result of different carrier-phase tracking alignments implemented in different receivers and the resulting biases

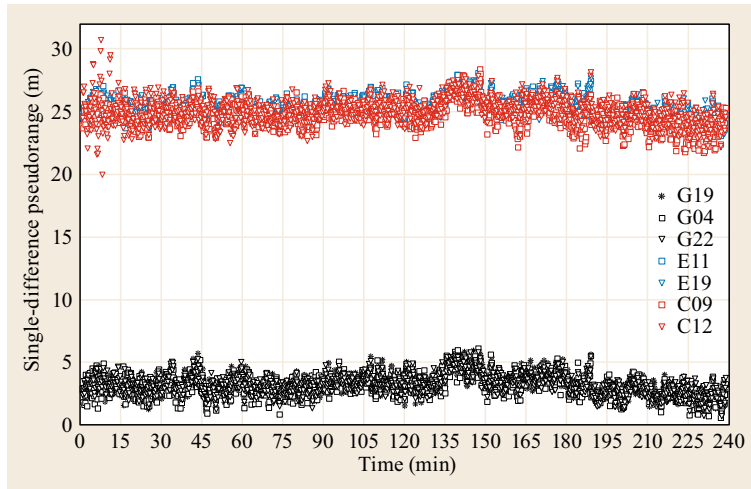


Fig. 20.13 Single differences of pseudorange observations for GPS (black), Galileo (blue), and BeiDou (red) satellites on a zero-baseline over a time interval of 4 h

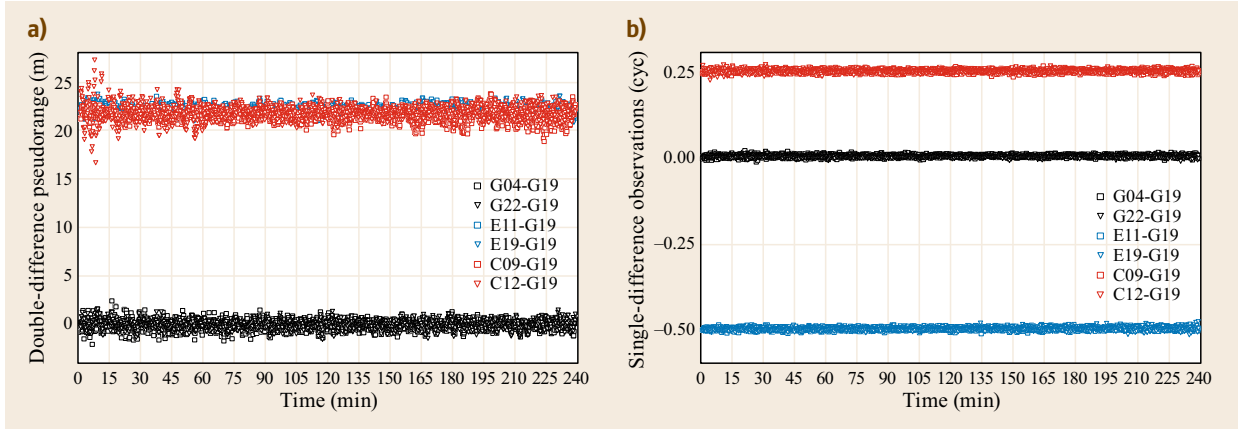


Fig. 20.14a,b Double differences of pseudorange (a) and carrier-phase (b) observations for GPS (black), Galileo (blue), and BeiDou (red) satellites on a zero-baseline over a time interval of 4 h. Note the different scales of the ordinates for pseudorange (m) and carrier-phase (cycles)

must be corrected for or estimated in order to enable inter-constellation ambiguity resolution [20.34, 35].

The examples for double differences of observations on a zero-baseline have demonstrated that all nuisance parameters except for differential biases, noise,

and multipath can be eliminated from the observations. This makes zero-baseline configurations particularly useful for receiver or signal tests, for example, for the characterization of elevation-dependent noise or receiver and satellite biases.

20.4 Pseudorange Filtering

The combination of GNSS observations discussed in the preceding sections typically either eliminated certain nuisance parameters (e.g., ionospheric path delays), isolated certain errors or effects (e.g., multipath or ionospheric delays), and make them accessible to the analysis.

A special type of combination is given by the joint use of pseudorange and carrier-phase measurements to obtain pseudorange observations with a reduced noise level. As discussed in Chap. 22 in full detail, such carrier-phase-adjusted pseudoranges can be obtained from a series of pseudorange and phase observations using either a batch least squares estimator or a recursive filter.

A widespread implementation of this concept is the Hatch filter, which uses time differences of carrier-phase observations to obtain smoothed pseudoranges. Hatch filters are commonly used inside GNSS receivers to reduce the raw pseudorange noise.

Even though the single-channel phase-based pseudorange *smoothing* does not achieve the same quality as the rigorous estimation of phase-adjusted pseudoranges (Sect. 22.3.1 and [20.36]), for certain applications these limitations can be accepted and the Hatch filter is an attractive alternative due to its computational simplicity.

For the derivation of the Hatch filter, time differences of the pseudorange (20.2) and carrier-phase (20.4) observation are considered

$$\begin{aligned}\partial p_{r,j}^s &= \partial P_{r,IF}^s + \partial I_{r,j}^s + \partial e_{r,j}^s, \\ \partial \varphi_{r,j}^s &= \partial P_{r,IF}^s - \partial I_{r,j}^s + \partial e_{r,j}^s.\end{aligned}\quad (20.76)$$

Note that several terms have been omitted in the above equations. Based on the assumption that no cycle-slips occur, the constant carrier-phase ambiguity terms drop out in the time difference. Also pseudorange and carrier-phase biases are completely eliminated. The change in carrier-phase wind-up as well as group delay and phase-center variation is assumed to be negligibly small over the time interval.

If the change in the ionospheric delay is neglected, a predicted pseudorange at epoch t_i can be computed based on the observed pseudorange at the previous epoch t_i and the carrier-phase difference $\partial \varphi_{r,j}^s$

$$p_{r,j}^s(t_i) = p_{r,j}^s(t_{i-1}) + \partial \varphi_{r,j}^s. \quad (20.77)$$

The prediction based on the phase difference has low noise since $\partial e_{r,j}^s \ll \partial e_{r,j}^s$. This advantage can be exploited in a recursive filter, which produces a smoothed

pseudorange $\tilde{p}_{r,j}^s(t_i)$ based on the weighted average of the observed pseudorange $p_{r,j}^s(t_i)$ and the prediction from (20.77) using the smoothed value from the previous epoch

$$\tilde{p}_{r,j}^s(t_i) = w p_{r,j}^s(t_i) + (1 - w) (\tilde{p}_{r,j}^s(t_{i-1}) + \partial \varphi_{r,j}^s). \quad (20.78)$$

In the traditional formulation of the Hatch filter, the weight w is set to a constant value of $w = \tau / \tau_s$, where τ is the data rate of the observations and τ_s is the filter smoothing time constant. The filter is initialized with $\tilde{p}_{r,j}^s(t_0) = p_{r,j}^s(t_0)$ [20.37]. Other formulations use a time-dependent weighting factor, for example, starting with $w = 1$ when the filter is initialized and reducing the weight by a factor of τ / τ_s step-by-step with every new epoch until $w = 0$. After the time interval τ_s , the weight factor remains zero and the smoothed pseudoranges are only based on the carrier-phase predictions [20.38]. In both cases, a cycle-slip destroys the continuity of the prediction and requires the filter to be initialized again.

The Hatch filter is an efficient yet simple measure to reduce the noise of pseudorange measurements and improve the position solution based on the smoothed observations. Its simplicity comes with a disadvantage, though: due to the opposite sign of the ionospheric term in (20.76), the smoothed pseudorange may diverge significantly from original pseudorange, when the ionosphere is active and non-negligibly varying during the smoothing interval. A similar reasoning also holds true for the nonstochastic multipath errors. Thus, when selecting the weight for the Hatch filter, care must be taken to choose a value which yields sufficiently smoothed observations, but prevents significant errors due to code-carrier divergence.

The adverse effects of the ionosphere can be eliminated when dual-frequency measurements are available. One option is to replace the single-frequency carrier-phase time difference in (20.78) with a ionosphere-corrected carrier-phase $\partial \tilde{\varphi}_r^s$ based on (20.37)

$$\begin{aligned} \partial \tilde{\varphi}_r^s &= \partial \varphi_{r,A}^s + 2 \partial I_{r,A}^s \\ &= \partial \varphi_{r,A}^s + 2 \frac{f_B^2}{f_A^2 - f_B^2} (\partial \varphi_{r,A}^s - \partial \varphi_{r,B}^s), \end{aligned} \quad (20.79)$$

where f_A and f_B are the corresponding frequencies to the differential carrier-phase observations $\partial \varphi_{r,A}^s$ and $\partial \varphi_{r,B}^s$. The ionosphere-corrected carrier-phase has identical ionospheric delay variation as the time difference of pseudorange and the instantaneous ionospheric delay is retained in the smoothed pseudorange. This approach is referred to in literature as divergence-free smoothing [20.39].

The second option is to use a ionosphere-free combination of pseudorange and carrier-phase observations according to (20.41) and (20.42) as input for the Hatch filter. This method is often called ionosphere-free smoothing. The result is then free of ionospheric effects, but affected by the increased noise of the ionosphere-free combination [20.39].

In order to visualize the effects of filter divergence, simulated pseudorange and carrier-phase measurements at a data rate of 1 Hz with realistically modeled ionospheric error and measurement noise have been processed with different smoothing intervals in a Hatch filter with constant weight factor. The results are depicted in Fig. 20.15. The errors of the raw pseudoranges without smoothing compared to the true pseudoranges are plotted as black diamonds. Note that only the receiver noise appears here, since both raw and true

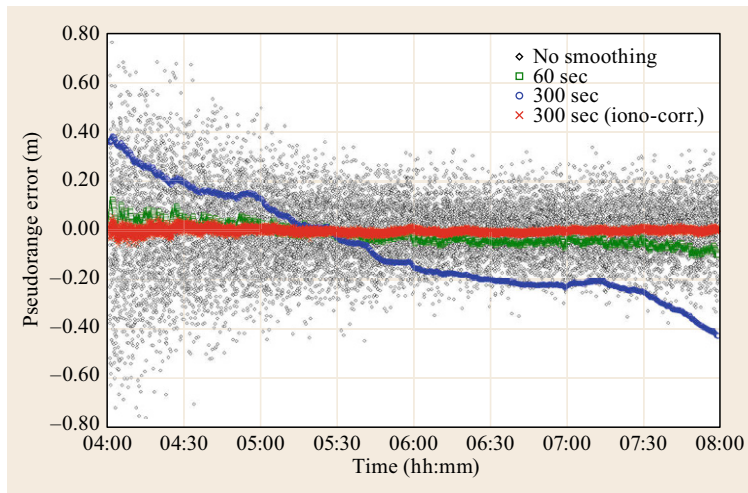


Fig. 20.15 Results for smoothed pseudorange observations over time from different Hatch filters based on simulated measurements. The plot shows the errors of the original observations (*black diamonds*), the results of single-frequency Hatch filters with smoothing intervals of 60 s (*green squares*) and 300 s (*blue circles*), and a divergence-free Hatch filter with 300 s smoothing interval (*red crosses*)

pseudoranges contain the ionospheric delay. The single-frequency Hatch filter with 60 s smoothing interval significantly reduces the noise, but small systematic effects are already visible. For a smoothing interval of 300 s, the filtered measurements are smoother, but large systematic pseudorange errors are visible. Making use of a second carrier-phase measurement in a divergence-free Hatch filter completely eliminated the systematic errors despite the same smoothing interval. However, note the slightly increased noise in the filtered pseudorange.

It should finally be mentioned that Hatch filtering cannot only be applied to observations of a single receiver, but also to single or double differences of observations. These techniques are relevant, for example, in case of differential GPS (DGPS) for aircraft landing. Even though most of the ionospheric delay is already canceled out through differential-processing, which may in general allow for longer smoothing intervals, the divergence of the Hatch filter can still become significant during anomalous ionospheric conditions like ionospheric storms [20.39, 40].

References

- 20.1 P. Enge, P. Misra: *Global Positioning System: Signals, Measurements, and Performance*, 2nd edn. (Ganga-Jamuna, Lincoln 2006)
- 20.2 P. Henkel, C. Günther: Reliable integer ambiguity resolution: Multi-frequency code carrier linear combinations and statistical a priori knowledge of attitude, *Navigation* **59**(1), 61–75 (2012)
- 20.3 M. Cocard, S. Bourgon, O. Kamali, P. Collins: A systematic investigation of optimal carrier-phase combinations for modernized triple-frequency GPS, *J. Geod.* **82**(9), 555–564 (2008)
- 20.4 X. Zhang, X. He: BDS triple-frequency carrier-phase linear combination models and their characteristics, *Sci. China Earth Sci.* **58**(6), 896–905 (2015)
- 20.5 M. Cocard, A. Geiger: Systematic search for all possible widelanes, *Proc. 6th Int. Geod. Symp. Satell. Position.*, Columbus (1992) pp. 312–318
- 20.6 P. Collins: An overview of inter-frequency carrier phase combinations (1999) <http://gauss.gge.unb.ca/papers.pdf/L1L2combinations.collins.pdf>
- 20.7 J. Jung: High integrity carrier phase navigation for future LAAS using multiple civilian GPS signals, *Proc. ION GPS 1999*, Nashville (ION, Virginia 1999) pp. 727–736
- 20.8 B. Forssell, M. Martin-Neira, R.A. Harris: Carrier phase ambiguity resolution in GNSS-2, *Proc. ION GPS 1997*, Kansas City (ION, Virginia 1997) pp. 1727–1736
- 20.9 U. Vollath, S. Birnbach, H. Landau: Analysis of three-carrier ambiguity resolution (TCAR) technique for precise relative positioning in GNSS-2, *Proc. ION GPS 1998*, Nashville (ION, Virginia 1998) pp. 417–426
- 20.10 R. Hatch, J. Jung, P. Enge, B. Pervan: Civilian GPS: The benefit of three frequencies, *GPS Solutions* **3**(4), 1–9 (2000)
- 20.11 J. Jung, P. Enge, B. Pervan: Optimization of cascade integer resolution with three civil GPS frequencies, *Proc. ION GPS 2000*, Salt Lake City (ION, Virginia 2000) pp. 2191–2200
- 20.12 P.J.G. Teunissen, P. Joosten, C. Tiberius: A comparison of TCAR, CIR and LAMBDA GNSS ambiguity resolution, *Proc. ION GPS 2002*, Portland (ION, Virginia 2002) pp. 2799–2808
- 20.13 S. Ji, W. Chen, C. Zhao, X. Ding, Y. Chen: Single epoch ambiguity resolution for Galileo with the CAR and LAMBDA methods, *GPS Solutions* **11**(4), 259–268 (2007)
- 20.14 K. O'Keefe, M. Petovello, W. Cao, G. Lachapelle, E. Guyader: Comparing multicarrier ambiguity resolution methods for geometry-based GPS and Galileo relative positioning and their application to low Earth orbiting satellite attitude determination, *Int. J. Navig. Obs.*, 592073 (2009) doi:10.1155/2009/592073
- 20.15 W.G. Melbourne: The case for ranging in GPS based geodetic systems, *Proc. 1st Int. Symp. Precise Position. Glob. Position. Syst.*, Rockville, ed. by C. Goad (NOAA, Washington DC 1985) pp. 373–386
- 20.16 G. Wübbena: Software developments for geodetic positioning with GPS using TI 4100 code and carrier measurements, *Proc. 1st Int. Symp. Precise Position. Glob. Position. Syst.*, Rockville, ed. by C. Goad (NOAA, Washington DC 1985) pp. 403–412
- 20.17 G. Blewitt: An automatic editing algorithm for GPS data, *Geophys. Res. Lett.* **17**(3), 199–202 (1990)
- 20.18 S. Han, C. Rizos: The impact of two additional civilian GPS frequencies on ambiguity resolution strategies, *Proc. ION Annu. Meet. 1999* (ION, Virginia 1999) pp. 315–321
- 20.19 P. Henkel, C. Günther: Three frequency linear combinations for Galileo, *Proc. 4th Workshop Position., Navig. Commun.*, Hannover (2007) pp. 239–245
- 20.20 T.P. Yunck: Coping with the atmosphere and ionosphere in precise satellite and ground positioning. In: *Environmental Effects on Spacecraft Positioning and Trajectories*, ed. by A.V. Jones (AGU, Washington DC 1992) pp. 1–16
- 20.21 D. Odijk: Ionosphere-free phase combinations for modernized GPS, *J. Surv. Eng.* **129**(4), 165–173 (2003)
- 20.22 P.J.G. Teunissen, D. Odijk: Rank-defect integer estimation and phase-only modernized GPS ambiguity resolution, *J. Geod.* **76**(9/10), 523–535 (2003)
- 20.23 Z. Wang, Y. Wu, K. Zhang, Y. Meng: Triple-frequency method for high-order ionospheric refractive error modelling in GPS modernization, *J. Glob. Position. Syst.* **4**(1/2), 291–295 (2005)

- 20.24 M. Hernández-Pajares, Á. Aragón-Ángel, P. De-fraigne, N. Bergeot, R. Prieto-Cerdeira, A. Garcea-Rigo: Distribution and mitigation of higher-order ionospheric effects on precise GNSS processing, *J. Geophys. Res. Solid Earth* **119**(4), 3823–3837 (2014)
- 20.25 T. Richert, N. El-Sheimy: Optimal linear combinations of triple frequency carrier phase data from future global navigation satellite systems, *GPS Solutions* **11**(1), 11–19 (2007)
- 20.26 Y. Feng: GNSS three carrier ambiguity resolution using ionosphere-reduced virtual signals, *J. Geod.* **82**(12), 847–862 (2008)
- 20.27 C. Kee, B. Parkinson: Calibration of multipath errors on GPS pseudorange measurements, *Proc. ION GPS 1994*, Salt Lake City (ION, Virginia 1994) pp. 353–362
- 20.28 P.D. Groves, Z. Jiang, M. Rudi, P. Strode: A portfolio approach to NLOS and multipath mitigation in dense urban areas, *Proc. ION GNSS+ 2013*, Nashville (ION, Virginia 2013) pp. 3231–3247
- 20.29 A. Simsky: Three's the charm: Triple-frequency combinations in future GNSS, *Inside GNSS* **1**(5), 38–41 (2006)
- 20.30 O. Montenbruck, A. Hauschild, P. Steigenberger, R.B. Langley: Three's the challenge: A close look at GPS SVN62 triple-frequency signal combinations finds carrier-phase variations on the new L5, *GPS World* **21**(8), 8–19 (2010)
- 20.31 O. Montenbruck, U. Hugentobler, R. Dach, P. Steigenberger, A. Hauschild: Apparent clock variations of the block IIF-1 (SVN62) GPS satellite, *GPS Solutions* **16**(3), 303–313 (2012)
- 20.32 J.T. Wu, S.C. Wu, G.A. Hajj, W.I. Bertiger, S.M. Lichten: Effects of antenna orientation on GPS carrier phase, *Manuscr. Geod.* **18**(2), 91–98 (1993)
- 20.33 G. Blewitt: Basics of the GPS technique: Observation equations. In: *Geodetic Applications of GPS*, ed. by B. Jonsson (Swedish Land Survey, Gävle 1997)
- 20.34 D. Odijk, P.J.G. Teunissen: Characterization of between-receiver GPS-Galileo inter-system biases and their effect on mixed ambiguity resolution, *GPS Solutions* **17**(4), 521–533 (2013)
- 20.35 N. Nadarajah, P.J.G. Teunissen, J.-M. Sleewaegen, O. Montenbruck: The mixed-receiver BeiDou inter-satellite-type bias and its impact on RTK positioning, *GPS Solutions* **19**(3), 357–368 (2015)
- 20.36 P.J.G. Teunissen: The GPS phase-adjusted pseudorange, *Proc. 2nd Int. Workshop High Precis. Navig.*, Stuttgart/Freudenstadt, ed. by K. Linkwitz, U. Hangleiter (Dümmler, Bonn 1991) pp. 115–125
- 20.37 L. Zhao, L. Li, X. Zhao: An adaptive Hatch filter to minimize the effects of ionosphere and multipath for GPS single point positioning, *Proc. IEEE Int. Conf. Mechatron. Autom.*, Changchun (2009) pp. 4167–4172
- 20.38 B. Hoffmann-Wellenhof, H. Lichtenegger, E. Wasle: *GNSS-Global Navigation Satellite Systems* (Springer, Wien, New York 2008)
- 20.39 G.A. McGraw, R.S.Y. Young: Dual frequency smoothing DGPS performance evaluation studies, *Proc. ION NTM 2005*, San Diego (ION, Virginia 2005) pp. 170–181
- 20.40 P. Hwang, G. McGraw, J. Bader: Enhanced differential GPS carrier-smoothed code processing using dual-frequency measurements, *Navigation* **46**(2), 127–137 (1999)

Positioning Model

Dennis Odijk

The focus of this chapter is on the models for positioning. Since the global navigation satellite system (GNSS) observation equations are nonlinear in the position coordinates, the chapter is started with a section on the linearization of the observation equations for pseudorange (code) and carrier-phase. After that, absolute (point) positioning models are discussed, starting with the code-based single point positioning (SPP) model, followed by the model for precise point positioning (PPP), based on code and phase. The relative positioning models can be distinguished into code-dominated (differential GNSS or DGNSS) models and phase-dominated (real-time kinematic or RTK) models. For the latter type of models, a general multifrequency undifferenced model is presented, which may form the basis of both relative network model and the (absolute) model that enables PPP users to perform integer ambiguity resolution (*PPP-RTK*). After that the link is made between the undifferenced model and the single and double differenced versions of the positioning model and an overview is given of the various positioning concepts.

21.1	Nonlinear Observation Equations	606	21.3	Point Positioning Models	612
21.1.1	Single-GNSS Observation Equations	606	21.3.1	Computation of the Satellite Clocks and Hardware Code (Group) Delays	613
21.1.2	Multi-GNSS Observation Equations	607	21.3.2	Some Remarks on the TGDs/DCBs	615
21.2	Linearization of the Observation Equations	609	21.3.3	Computation/Estimation of the Atmospheric Errors	615
21.2.1	Linearizing the Receiver-Satellite Range	609	21.3.4	Single-Constellation SPP Model	615
21.2.2	Linearized Observation Equations	612	21.3.5	Multiconstellation SPP Model	617
			21.3.6	Precision and DOP	618
			21.3.7	PPP Model	619
			21.4	Relative Positioning Models	623
			21.4.1	Principle of DGNSS and (PPP-)RTK	623
			21.4.2	Impact of Orbit Errors	625
			21.4.3	Ionosphere-Fixed/Weighted/Float Models	625
			21.4.4	Undifferenced Relative Positioning Models	625
			21.4.5	PPP-RTK Models	627
			21.4.6	Link Between PPP-RTK and PPP	630
			21.5	Differenced Positioning Models	631
			21.5.1	Single Differencing	631
			21.5.2	Double and Triple Differencing	632
			21.5.3	Redundancy of the Differenced Models	633
			21.6	The Positioning Concepts Related	633
			21.6.1	Global Positioning: SPP/PPP	633
			21.6.2	Regional Positioning: Network DGNSS/RTK	634
			21.6.3	Local Positioning: Single-Baseline DGNSS/RTK	634
			21.6.4	Global/Regional Positioning: PPP-RTK	634
			21.6.5	Accuracy of the Positioning Concepts	635
			References		635

21.1 Nonlinear Observation Equations

In this section, the nonlinear observations equations for code and phase are reviewed, first for a single global or regional navigation satellite system (GNSS or RNSS) and after that the observation equations in case the GNSS receiver tracks observations from more than one constellation (the multiconstellation case).

21.1.1 Single-GNSS Observation Equations

Recall from Chap. 19, the nonlinear observation equations for $j = 1, \dots, f_s$ frequencies of a certain GNSS constellation S, at time of observation t in the GNSS system time. Then the observation equation for the code or pseudorange from satellite s tracked by receiver r at epoch t can be given as

$$\begin{aligned} p_{r,j}^s(t) = & \rho_r^s(t, t - \tau_r^s) + T_r^s(t) \\ & + c [dt_r(t) + d_{r,j}^s(t) + \Delta d_{r,j}^s(t)] \\ & - c [dt^s(t - \tau_r^s) - d_j^s(t - \tau_r^s)] \\ & + \mu_j^s I_r^s(t) + e_{r,j}^s(t), \end{aligned} \quad (21.1)$$

while the carrier-phase observation equation reads

$$\begin{aligned} \varphi_{r,j}^s(t) = & \rho_r^s(t, t - \tau_r^s) + T_r^s(t) \\ & + c [dt_r(t) + \delta_{r,j}^s(t) + \Delta \delta_{r,j}^s(t)] \\ & - c [dt^s(t - \tau_r^s) - \delta_j^s(t - \tau_r^s)] \\ & - \mu_j^s I_r^s(t) + \lambda_j^s N_{r,j}^s + \varepsilon_{r,j}^s(t). \end{aligned} \quad (21.2)$$

The constellation identifier S is chosen in agreement with the RINEX convention (Annex A *Data Formats*) such that $S \in \{G, R, E, C, J, I, \dots\}$, for GPS, GLONASS, Galileo, BeiDou (BDS), QZSS, IRNSS, etc. The notation in (21.1) and (21.2) is as follows

$p_{r,j}^s$	Code/pseudorange observable (m)
$\varphi_{r,j}^s$	Carrier-phase observable (m)
ρ_r^s	Receiver-satellite range (m)
τ_r^s	Signal travel time (s)
T_r^s	Tropospheric delay (m)
c	Velocity of light (m/s)
dt_r	Receiver clock error (s)
dt^s	Satellite clock error (s)
$d_{r,j}^s$	Receiver code hardware bias (s)
$\delta_{r,j}^s$	Receiver phase hardware bias (s)
$\Delta d_{r,j}^s$	Code interchannel bias (s)
$\Delta \delta_{r,j}^s$	Phase interchannel bias (s)
d_j^s	Satellite code hardware bias (s)
δ_j^s	Satellite-phase hardware bias (s)
μ_j^s	Ionospheric coefficient
I_r^s	Ionospheric delay (m)

λ_j^s	Wavelength (m)
$N_{r,j}^s$	Carrier-phase ambiguity (cyc)
$e_{r,j}^s$	Random code noise (m)
$\varepsilon_{r,j}^s$	Random carrier-phase noise (m).

The receiver hardware biases in (21.1) and (21.2), denoted as $d_{r,j}^s(t)$ and $\delta_{r,j}^s(t)$, are in principle different for each constellation (that is why they have a constellation index S), even when the signals are tracked on frequency bands that overlap between the constellations [21.1], as for example GPS L1 and Galileo E1. These hardware biases are caused by various reasons, including analog group delays in the frontend and digital delays. The correlation process in the receiver affects the resulting delays as well [21.2]. The difference in receiver hardware biases between signals of different constellations is referred to as *intersystem bias* (ISB) [21.3–5].

For constellations that transmit signals based on the *frequency division multiple access* (FDMA) technology (Chap. 4), the frequency is different per channel. In case of the GLONASS FDMA signals, the L1 frequency equals $f_1^{\text{RS}} = 1602 + \kappa^s(9/16)$ MHz (Chap. 8), where κ^s denotes the channel number that can take on the following integers: $\kappa^s \in \{-7, -6, \dots, +5, +6\}$. The GLONASS L2 frequency equals $f_2^{\text{RS}} = 1246 + \kappa^s(7/16)$ MHz. In case of FDMA signals, the code and phase observation equations are also contaminated by *interchannel biases* (ICBs), denoted by $\Delta d_{r,j}^s(t)$ and $\Delta \delta_{r,j}^s(t)$. For signals that are based on the *code division multiple access* (CDMA) technology, the frequency is identical for all channels and the satellite index can thus be omitted. Also no ICBs show up for CDMA signals: $\Delta d_{r,j}^s(t) = 0$ and $\Delta \delta_{r,j}^s(t) = 0$.

It is emphasized that both satellite code bias and phase bias in (21.1) and (21.2), denoted as $d_j^s(t - \tau_r^s)$ and $\delta_j^s(t - \tau_r^s)$, respectively, are, like the receiver hardware biases, considered as additive parameters, that is, they have a (net) plus sign in the observation equations (whereas the satellite clock has a minus sign). The reason for doing so is to be consistent with the convention adopted by the International GNSS Service (IGS) [21.6].

In the above observation equations, it has been implicitly assumed that (frequency-dependent) offsets between the center of mass of the satellite and the satellite antenna reference point, as well as (frequency-dependent) offsets between the receiver antenna reference point and the receiver's point to be determined by positioning, can be taken into account through dedicated terms on the right-hand side of the observation equations. However, within the following discussion,

the terms are not further considered for the ease of notation, and the receiver-satellite range $\rho_r^s(t, t - \tau_r^s)$ is taken common for both code and phase, as well as for all frequencies.

In the code and phase-observation equations, the dispersive (first-order) ionospheric delays are mapped to one frequency, that is, $I_{r,j}^s(t) = \mu_j^s I_r^s(t)$. Usually $I_r^s(t)$ denotes the ionospheric delay for the first frequency, such that the frequency-dependent ionospheric coefficient is defined as follows

$$\mu_j^s = \left(\frac{\lambda_j^s}{\lambda_1^s} \right)^2 = \left(\frac{f_1^s}{f_j^s} \right)^2. \quad (21.3)$$

From this definition, it follows that for the first frequency ($j = 1$), the ionospheric coefficient equals $\mu_1^s = 1$. In case of GLONASS FDMA signals, although the frequencies differ per channel, their (squared) L1-L2 ratio is *independent* of the channel, since in case of dual-frequency GLONASS

$$\mu_2^R = \left(\frac{\lambda_2^{Rs}}{\lambda_1^{Rs}} \right)^2 = \left(\frac{f_1^{Rs}}{f_2^{Rs}} \right)^2 = \left(\frac{9}{7} \right)^2. \quad (21.4)$$

Note that for dual-frequency GPS $\mu_2^G = (77/60)^2$.

Effects that have not been accounted for in (21.1) and (21.2) are, among others, phase-center offsets and variations, phase wind-up (phase only), relativistic effects, a-priori tropospheric model, etc. (Chap. 19). For the present chapter they are dropped from the observation equations to ease the notation and to focus on the illustration of the basic positioning concepts.

21.1.2 Multi-GNSS Observation Equations

Constellation-Specific Time Frames

To derive the observation equations for multiple constellations, for simplicity it is assumed that a multi-GNSS receiver tracks data of two constellations, denoted as A and B. If the observations of system A are collected at receiver time t_r (this is the time tag in the RINEX observation file), this (measured) receiver time deviates from the (unknown) system time of the first constellation t^A by means of a receiver clock error dt_r

$$t_r(t^A) = t^A + dt_r(t^A). \quad (21.5)$$

Note that for reasons of simplicity, we have ignored effects due to receiver hardware delays and other errors like receiver noise and multipath in above expression. Observations of system B that are collected at the same receiver time t_r use different physical clocks to realize their own GNSS system time [21.7]. However, they can

be expressed as function of the receiver clock error in the system time of A

$$t_r(t^B) = t^B + dt_r(t^B) = t^B + dt_r(t^A) - t^{AB}, \quad (21.6)$$

with $t^{AB} = t^B - t^A$, the *system time offset* (thus: $t_r(t^A) = t_r(t^B)$, see also Fig. 21.1). In case the first constellation is GPS and the second is Galileo, this offset is also known as GPS-to-Galileo time offset (GGTO) [21.8]. The time of transmission at a satellite of constellation A, which is denoted using superscript s , reads, ignoring satellite hardware delays

$$t^s(t^A - \tau_r^s) = t^A - \tau_r^s + dt^s(t^A - \tau_r^s). \quad (21.7)$$

For a satellite of constellation B, denoted by superscript q , it reads

$$t^q(t^B - \tau_r^q) = t^B - \tau_r^q + dt^q(t^B - \tau_r^q). \quad (21.8)$$

Converted to pseudoranges this yields for the two constellations, now including atmospheric delays, hardware delay parameters and noise terms

$$\begin{aligned} p_{r,j}^s(t^A) &= c [t_r(t^A) - t^s(t^A - \tau_r^s)] \\ &= \rho_r^s(t^A, t^A - \tau_r^s) + T_r^s(t^A) \\ &\quad + c [dt_r(t^A) + d_{r,j}^A(t^A) + \Delta d_{r,j}^s(t^A)] \\ &\quad - c [dt^s(t^A - \tau_r^s) - d_j^s(t^A - \tau_r^s)] \\ &\quad + \mu_j^A I_r^s(t^A) + e_{r,j}^s(t^A), \\ p_{r,j}^q(t^B) &= c [t_r(t^B) - t^q(t^B - \tau_r^q)] \\ &= \rho_r^q(t^B, t^B - \tau_r^q) + T_r^q(t^B) \\ &\quad + c [dt_r(t^A) - t^{AB} + d_{r,j}^B(t^B) + \Delta d_{r,j}^q(t^B)] \\ &\quad - c [dt^q(t^B - \tau_r^q) - d_j^q(t^B - \tau_r^q)] \\ &\quad + \mu_j^B I_r^q(t^B) + e_{r,j}^q(t^B). \end{aligned} \quad (21.9)$$

Note that instead of the receiver clock in the time system of B, we have used the receiver clock in the

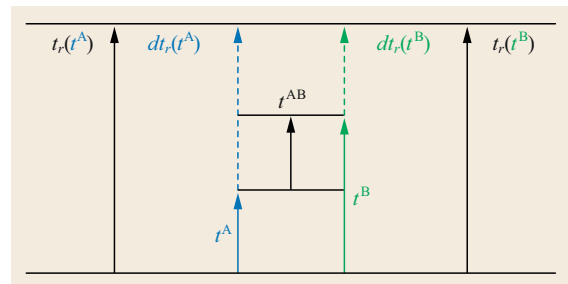


Fig. 21.1 Relation between time frames, receiver time, receiver clock error, and time offset of constellations A and B

time system of A, together with the system time offset, making use of (21.6), from which follows that $dt_r(t^B) = dt_r(t^A) - t^{AB}$. We can do something similar for the receiver hardware bias of the observations of constellations B, making use of the following definition of the ISB

$$\text{ISB}_{r,j}^{AB}(t^A, t^B) = [d_{r,j}^B(t^B) - d_{r,j}^A(t^A)] + [\Delta d_{r,j}^q(t^B) - \Delta d_{r,j}^s(t^A)]. \quad (21.10)$$

It is remarked that the interchannel terms only appear in case one of the constellations is based on FDMA. In that case, the ISB becomes satellite dependent; otherwise it is receiver dependent. Based on the ISB reparameterization, the code-observation equation for constellation B can be rewritten as

$$\begin{aligned} p_{r,j}^q(t^B) &= \rho_r^q(t^B, t^B - \tau_r^q) + T_r^q(t^B) \\ &+ c [dt_r(t^A) + d_{r,j}^A(t^A) + \Delta d_{r,j}^s(t^A)] \\ &+ c [\text{ISB}_{r,j}^{AB}(t^A, t^B) - t^{AB}] \\ &- c [dt_r^q(t^B - \tau_r^q) - d_j^q(t^B - \tau_r^q)] \\ &+ \mu_j^B I_r^q(t^B) + e_{r,j}^q(t^B). \end{aligned} \quad (21.11)$$

Compared to the code-observation equation of constellation A (see first equation in (21.9)), its counterpart for constellation B is now given as a function of the receiver clock, receiver hardware bias, as well as ICB (in case of GLONASS FDMA) of signals of constellation A. Also the ionospheric delays for the signals of constellation B can be expressed as ionospheric delays on the first frequency of constellation A by setting the ionospheric coefficient for B equal to $\mu_j^B = (\lambda_j^B / \lambda_1^A)^2$. For the phase-observation equation of constellation B a similar derivation can be made. Advantage of the formulation that involves an ISB parameter over the original formulation is that under certain conditions it is possible to *calibrate* the ISBs. When the ISB and also the system time offset are known, the observations of the two constellations can be processed *as if* they correspond to one system.

The code observation equation (21.11) is written as a function of the time stamps in two different systems, that is, t^A and t^B . For most GNSS systems, the differences between the time systems are sufficiently small, such that they may be neglected for the evaluation of observables and parameters in (21.11). This also holds for the purpose of the evaluation of the times of transmission at the satellites (Sect. 21.4.1). For these purposes from now on, we will simply use a common t for the time stamps of different systems. However,

the system time offset that itself is present as parameter in (21.11), that is, t^{AB} , may *not* be ignored in the observation equations of the second constellation, since it is multiplied by the velocity of light. For example, the offset between GPS time (GPST) and Galileo system time (GST) can be several tens of nanoseconds or tens of meters (Chap. 9 or [21.9]). The offset between QZSS time (QZSST) and GPST is less than two meters [21.10]. The offset between IRNSS time (IRNSSST) and GPST can be up to 3 m [21.11]. The difference between GPST and GLONASS system time can be several hundreds of nanoseconds (equivalent to hundreds of meters). The intersystem time offsets are broadcast as part of the navigation messages [21.12, 13] such that a user can correct his observations. The offset between GPST and GLONASS system time are broadcast as part of the navigation message of the GLONASS-M satellites [21.14]. Alternatively, the user can treat the offset as unknown parameter in his processing.

Constellation-Specific Reference Frames

Besides that each GNSS constellation realizes its own system time, its broadcast satellite positions are defined in its own coordinate system, see Table 21.1 for an overview. In order to solve multiconstellation positioning models, all satellite positions need to be defined in one common reference frame. Otherwise, transformation parameters need to be estimated together with the other model parameters. Although the realization of the reference frames depends on the full deployment of the ground-station network of new constellations, the differences are expected to be small, as they are all realizations of the International Terrestrial Reference System (ITRS). WGS84 coincides with the ITRF at the level of a few centimeters (Chap. 2). The difference between GTRF and WGS84 is aimed at the level of 3 cm [21.15], while the offset between JGS and WGS84 is expected to be less than 2 cm [21.10]. IRNSS uses WGS84 as its coordinate system [21.16]. The difference between CGCS2000 and WGS84 is at the level of a few centimeters. Also, the latest release PZ-90.11 of the GLONASS reference frame is consistent with ITRF at epoch 2011.0 at the centimeter level [21.17].

Table 21.1 Reference frames for GNSS–RNSS constellations

GNSS	Reference frame
GPS	World Geodetic System 1984 (WGS84)
GLONASS	Parametry Zemli 1990 (PZ-90)
Galileo	Galileo Terrestrial Reference Frame (GTRF)
BeiDou	China Geodetic Coordinate System 2000 (CGCS2000)
QZSS	Japanese Geodetic System (JGS)
IRNSS	World Geodetic System 1984 (WGS84)

For (multiconstellation) SPP (Sect. 21.3.5), of which the positioning accuracy is at the 10 m level, it is not needed to take the differences between the constellation-specific reference frames into account. Also for relative (short-baseline) positioning these differences do not have to be taken into account, as they are, similar to orbit errors, canceled out if the baseline is of re-

stricted length (Sect. 21.4.2). For PPP(-RTK) and long-baseline applications (Sect. 21.3.7 and Sect. 21.4) one can, however, not ignore these differences and they need to be accounted for. It should be noted that in case of *precise* IGS orbits, the satellite positions of the different constellations are all defined with respect to one reference frame (i. e., ITRF; Chap. 33).

21.2 Linearization of the Observation Equations

Since the GNSS observation equations are nonlinear in the parameters of interest, that is, the receiver position coordinates, they need to be *linearized*. In this section, the focus is on the linearization of the receiver–satellite range.

21.2.1 Linearizing the Receiver–Satellite Range

The GNSS observation equations are nonlinear in both the receiver and satellite position. Using $t = t_r(t) - dt_r(t)$ (see (21.5)), the receiver–satellite range can be written as the following function

$$\begin{aligned}\rho_r^s(t, t - \tau_r^s) &= \|\mathbf{r}^s(t - \tau_r^s) - \mathbf{r}_r(t)\| \\ &= \|\mathbf{r}^s [t_r(t) - dt_r(t) - \tau_r^s] \\ &\quad - \mathbf{r}_r[t_r(t) - dt_r(t)]\|. \end{aligned} \quad (21.12)$$

The norm of a vector is defined as $\|\cdot\| = \sqrt{(\cdot)^\top (\cdot)}$, where $(\cdot)^\top$ denotes a transposed vector (or matrix). The satellite position vector reads $\mathbf{r}^s = [x^s, y^s, z^s]^\top$ and the receiver position vector $\mathbf{r}_r = [x_r, y_r, z_r]^\top$.

With the true GNSS time t unknown, according to *Taylor's theorem* (Chap. 22) the receiver–satellite range can be linearized with respect to the unknown receiver position $\mathbf{r}_r(t)$, satellite position $\mathbf{r}^s(t - \tau_r^s)$, and receiver clock error $dt_r(t)$, as follows

$$\rho_r^s(t, t - \tau_r^s) \doteq \rho_r^s(t, t - \tau_r^s)|_0 + \Delta \rho_r^s(t, t - \tau_r^s). \quad (21.13)$$

The incremental receiver–satellite range is computed as

$$\begin{aligned}\Delta \rho_r^s(t, t - \tau_r^s) &= [\partial_{\mathbf{r}_r(t)} \rho_r^s(t, t - \tau_r^s)|_0]^\top \Delta \mathbf{r}_r(t) \\ &\quad + [\partial_{\mathbf{r}^s(t - \tau_r^s)} \rho_r^s(t, t - \tau_r^s)|_0]^\top \Delta \mathbf{r}^s(t - \tau_r^s) \\ &\quad + \partial_{dt_r(t)} \rho_r^s(t, t - \tau_r^s)|_0 \Delta dt_r(t). \end{aligned} \quad (21.14)$$

Here the incremental parameters are denoted as $\Delta(\cdot) = (\cdot) - (\cdot)|_0$, where (\cdot) denotes the original parameter and

$(\cdot)|_0$ its *approximate value*. For all positioning models discussed in this chapter it is assumed that the *satellite positions* are *known*, and need not be estimated in the positioning model. This implies that $\Delta \mathbf{r}^s(t - \tau_r^s) = 0$. Knowing the satellite positions means that they are computed either from the broadcast ephemeris transmitted in the navigation message, or from the more precise ephemeris made available by the IGS (Chap. 33).

The derivative of the range with respect to the receiver position can be given as

$$\partial_{\mathbf{r}_r(t)} \rho_r^s(t, t - \tau_r^s)|_0 = - \frac{[\mathbf{r}^s(t - \tau_{r,0}^s) - \mathbf{r}_{r,0}(t)]}{\underbrace{\|\mathbf{r}^s(t - \tau_{r,0}^s) - \mathbf{r}_{r,0}(t)\|}_{\mathbf{e}_{r,0}^s(t)}}, \quad (21.15)$$

with $\mathbf{e}_{r,0}^s(t)$ denoting the *line-of-sight* (LOS) vector of unit length. The derivative of the range with respect to the receiver clock error can be computed as

$$\partial_{dt_r(t)} \rho_r^s(t, t - \tau_r^s)|_0 = \underbrace{\frac{\partial \rho_r^s(t, t - \tau_r^s)}{\partial t}}_{\dot{\rho}_{r,0}^s(t)} \bigg|_0 \underbrace{\frac{\partial t}{\partial dt_r(t)}}_{-1} \bigg|_0. \quad (21.16)$$

The time derivative of the receiver–satellite range that shows up here, denoted as $\dot{\rho}_{r,0}^s(t)$, is also referred to as *range rate*. The time derivative of the time itself with respect to the receiver clock error follows from $t = t_r(t) - dt_r(t)$. Summarizing, the linearized receiver–satellite range can be compactly presented as

$$\begin{aligned}\Delta \rho_r^s(t, t - \tau_r^s) &= -[\mathbf{e}_{r,0}^s(t)]^\top \Delta \mathbf{r}_r(t) \\ &\quad - \dot{\rho}_{r,0}^s(t) \Delta dt_r(t). \end{aligned} \quad (21.17)$$

The computation of the LOS vector $\mathbf{e}_{r,0}^s(t)$, as well as the receiver–satellite range $\rho_r^s(t, t - \tau_r^s)|_0$, requires the availability of the satellite position vector $\mathbf{r}^s(t - \tau_{r,0}^s)$.

Also the range rate $\dot{\rho}_{r,0}^s(t)$ needs to be computed. How this can be done is explained in the following subsections.

Computation of the Receiver–Satellite Ranges, Satellite Positions, and Line-of-Sight Vectors

In the computation of the partial derivative in (21.15), we need to evaluate the receiver–satellite range based on the known satellite position, as well as the approximate receiver position. The problem is that we do not know the propagation time $\tau_r^s = \rho_r^s(t, t - \tau_r^s)/c$, since it is itself a function of the unknown receiver–satellite range. In addition, the satellite position must be calculated at *transmission time*, since the satellite range can change as much as 60 m from the time the signal was transmitted, to the time the signal was received, approximately 0.07 s later. If the receiver time was used instead, the error in computed range could be tens of meters.

We follow the procedure for determining the travel time and computation of the receiver–satellite range as described in [21.18]. In (21.12), the receiver and satellite position are assumed to be defined in an Earth-centered inertial (ECI) coordinate system. However, we want to use these positions given in an Earth-centered-Earth-fixed (ECEF) coordinate system, such as WGS-84 in case of GPS. Rewriting the receiver–satellite range expression in ECEF coordinates yields [21.19]

$$\rho_r^s(t, t - \tau_r^s) = \|\mathbf{R}(t - \tau_r^s) \mathbf{r}_{\text{ECEF}}^s(t - \tau_r^s) - \mathbf{R}(t) \mathbf{r}_{r, \text{ECEF}}(t)\|. \quad (21.18)$$

The matrix $\mathbf{R}(T)$ describes the rotation from the ECEF to the ECI coordinate system, which reads (Chap. 2)

$$\mathbf{R}(T) = \begin{bmatrix} +\cos(\omega_\oplus T) & +\sin(\omega_\oplus T) & 0 \\ -\sin(\omega_\oplus T) & +\cos(\omega_\oplus T) & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (21.19)$$

Here T denotes the appropriate time argument and ω_\oplus the Earth's rotation rate (in rad/s) [21.18]. The inclusion of this rotation can be understood, since the Earth has rotated between the time of transmission and the time of reception of the signal. Using the property of rotations that $\mathbf{R}(t - \tau_r^s) = \mathbf{R}(t) \mathbf{R}(-\tau_r^s)$, the rotation at the time t can be taken outside the norm in (21.18), such that the receiver–satellite range expression becomes

$$\rho_r^s(t, t - \tau_r^s) = \|\mathbf{R}(-\tau_r^s) \mathbf{r}_{\text{ECEF}}^s(t - \tau_r^s) - \mathbf{r}_{r, \text{ECEF}}(t)\|. \quad (21.20)$$

The signal travel time τ_r^s as well as the satellite position at the time of transmission $\mathbf{r}_{\text{ECEF}}^s(t - \tau_r^s)$ can

now be determined using an *iterative procedure*. One starts with evaluating (21.20) with $\tau_r^s = 0$ and computes a new value of the travel time as $\tau_r^s = \rho_r^s(t, t)/c$. This value is used to compute a refined estimate of the receiver–satellite range. Usually three iterations are sufficient to get differences between the receiver–satellite ranges of the last two iterations within the order of 10^{-8} m [21.18]. Based on the iterated signal travel time and satellite position, the LOS vector is finally computed as

$$\mathbf{e}_{r, \text{ECEF}}^s(t) = \frac{\mathbf{R}(-\tau_r^s) \mathbf{r}_{\text{ECEF}}^s(t - \tau_r^s) - \mathbf{r}_{r, \text{ECEF}}(t)}{\|\mathbf{R}(-\tau_r^s) \mathbf{r}_{\text{ECEF}}^s(t - \tau_r^s) - \mathbf{r}_{r, \text{ECEF}}(t)\|}. \quad (21.21)$$

Note that $\mathbf{e}_{r, \text{ECEF}}^s(t) = \mathbf{R}(-t) \mathbf{e}_r^s(t)$. The above procedure requires approximate values for the receiver's position $\mathbf{r}_{r, \text{ECEF}}(t)$. These are available, provided that the above iterative procedure is integrated inside the Gauss–Newton iteration scheme (Chap. 22) to solve the nonlinear single point positioning (SPP) model (Sect. 21.3).

Computation of the Receiver–Satellite Range Rate

The derivative of the receiver–satellite range with respect to time can be computed from the projection of the relative satellite–receiver velocity onto the LOS vector [21.18]

$$\dot{\rho}_r^s(t) = \left[\frac{\partial (\mathbf{r}^s(t - \tau_r^s) - \mathbf{r}_r(t))}{\partial t} \right]^\top \mathbf{e}_r^s(t), \quad (21.22)$$

with \mathbf{r}^s and \mathbf{r}_r defined in the ECI frame. The time derivative of the satellite position can be computed as follows [21.20]

$$\begin{aligned} \frac{\partial \mathbf{r}^s(t - \tau_r^s)}{\partial t} &= \frac{\partial \mathbf{r}^s(t - \tau_r^s)}{\partial (t - \tau_r^s)} \frac{\partial (t - \tau_r^s)}{\partial t} \\ &= \dot{\mathbf{r}}^s(t - \tau_r^s) \frac{\partial \left(t - \frac{\rho_r^s(t, t - \tau_r^s)}{c} \right)}{\partial t} \\ &= \dot{\mathbf{r}}^s(t - \tau_r^s) \left[1 - \frac{\dot{\rho}_r^s(t)}{c} \right], \end{aligned} \quad (21.23)$$

where $\dot{\mathbf{r}}^s(t - \tau_r^s)$ denotes the satellite's velocity vector in the ECI frame. The receiver's velocity also appears in (21.22) and is denoted equivalently as $\frac{\partial \mathbf{r}_r(t)}{\partial t} = \dot{\mathbf{r}}_r(t)$, also in the ECI frame. Using this, we obtain the following expression for the range rate

$$\dot{\rho}_r^s(t) = \frac{[\dot{\mathbf{r}}^s(t - \tau_r^s) - \dot{\mathbf{r}}_r(t)]^\top \mathbf{e}_r^s(t)}{1 + \frac{1}{c} [\dot{\mathbf{r}}^s(t - \tau_r^s)]^\top \mathbf{e}_r^s(t)}. \quad (21.24)$$

This expression has also been derived in [21.21]. To evaluate it, the relative velocity of the GNSS satellite with respect to the receiver is required, as well as the relative geometry (line-of-sight). If we want to use velocity vectors in the ECEF frame, then the following relations hold between the vectors in the ECI and ECEF frames

$$\begin{aligned}\dot{\mathbf{r}}^s(t - \tau_r^s) &= \mathbf{R}(t - \tau_r^s) \dot{\mathbf{r}}_{\text{ECEF}}^s(t - \tau_r^s) \\ &\quad + \dot{\mathbf{R}}(t - \tau_r^s) \mathbf{r}_{\text{ECEF}}^s(t - \tau_r^s) \\ \dot{\mathbf{r}}_r(t) &= \mathbf{R}(t) \dot{\mathbf{r}}_{r,\text{ECEF}}(t) + \dot{\mathbf{R}}(t) \mathbf{r}_{r,\text{ECEF}}(t).\end{aligned}\quad (21.25)$$

The time derivative of the rotation matrix $\mathbf{R}(T)$ in (21.19) can be given as

$$\dot{\mathbf{R}}(T) = \omega_{\oplus} \begin{bmatrix} -\sin(\omega_{\oplus} T) & +\cos(\omega_{\oplus} T) & 0 \\ -\cos(\omega_{\oplus} T) & -\sin(\omega_{\oplus} T) & 0 \\ 0 & 0 & 0 \end{bmatrix}.\quad (21.26)$$

Note that the following property holds for the ECEF to ECI rotation matrix and its derivative

$$\dot{\mathbf{R}}(T)^\top \mathbf{R}(T) = \begin{bmatrix} 0 & -\omega_{\oplus} & 0 \\ \omega_{\oplus} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.\quad (21.27)$$

Using this result and with $\mathbf{e}_r^s(t) = \mathbf{R}(t) \mathbf{e}_{r,\text{ECEF}}^s(t)$, the range rate can be computed as follows from the vectors defined in ECEF

$$\dot{\rho}_r^s(t) = \frac{[\mathbf{v}^s(t - \tau_r^s) - \mathbf{v}_r(t)]^\top \mathbf{e}_{r,\text{ECEF}}^s(t)}{1 + \frac{1}{c} [\mathbf{v}^s(t - \tau_r^s)]^\top \mathbf{e}_{r,\text{ECEF}}^s(t)},\quad (21.28)$$

with

$$\begin{aligned}\mathbf{v}^s(t - \tau_r^s) &= \mathbf{R}(-\tau_r^s) \dot{\mathbf{r}}_{\text{ECEF}}^s(t - \tau_r^s) \\ &\quad + \boldsymbol{\omega} \times \dot{\mathbf{R}}(-\tau_r^s) \mathbf{r}_{\text{ECEF}}^s(t - \tau_r^s), \\ \mathbf{v}_r(t) &= \dot{\mathbf{r}}_{r,\text{ECEF}}(t) + \boldsymbol{\omega} \times \mathbf{r}_{r,\text{ECEF}}(t).\end{aligned}\quad (21.29)$$

Here use is made of the *vector cross product* \times and $\boldsymbol{\omega} = (0, 0, \omega_{\oplus})^\top$. In case of a static receiver on Earth, its velocity in the ECEF frame ($\dot{\mathbf{r}}_{r,\text{ECEF}}(t)$) equals zero, while for receivers in motion it can be assessed from positioning solutions at two epochs.

Local Coordinate System

Usually, the positioning model is solved for the position in an ECEF system, that is, $\mathbf{r}_{r,\text{ECEF}}(t)$. From now on, we will omit ECEF in the notation of the position and LOS vectors in this chapter, implicitly assuming they are with respect to the ECEF system. In addition, for the sake of interpretation of the position it is often convenient to work with another coordinate frame, that is, a *local* system which is centered at an assumed or approximate position, denoted as $\mathbf{r}_{r,0}$, of the point we would like to determine.

The x -axis of this local system is directed east, the y -axis directed north, and the z -axis is pointing upward and perpendicular to the local ellipsoidal surface (Fig. 21.2). Hence, this coordinate system is referred to as an *east–north–up* system. The coordinates of a point r in this system are denoted as $\mathbf{r}_{r,l} = [E_r, N_r, U_r]^\top$ (omitting the time stamp t). Such an east–north–up system is very practical for observers on or close to the Earth's surface, mainly for computation of dilution of precision (DOP) measures (Sect. 21.3.6), as well as for altitude-constrained (2-D) positioning. The ECEF coordinates

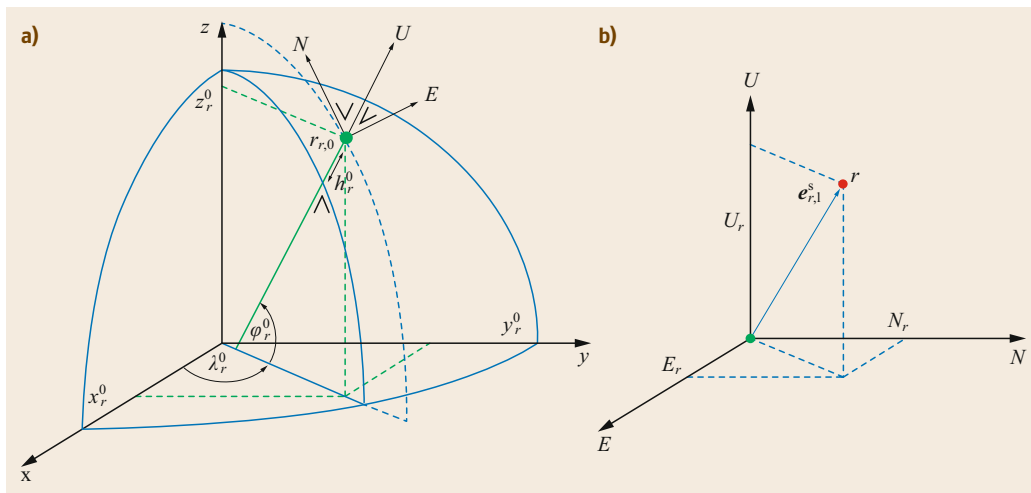


Fig. 21.2a,b East-north-up (ENU) local coordinate frame: **(a)** situated in the ECEF (xyz) frame with origin at $r, 0$ and **(b)** local ENU coordinates of point r

of point r , denoted as \mathbf{r}_r , can now be transformed to their east–north–up counterparts as follows

$$\mathbf{r}_{r,1} = \mathbf{R}_x \left(\frac{\pi}{2} - \varphi_r^0 \right) \mathbf{R}_z \left(\frac{\pi}{2} + \lambda_r^0 \right) [\mathbf{r}_r - \mathbf{r}_{r,0}] . \quad (21.30)$$

The product of the rotation matrices is elaborated as

$$\begin{aligned} & \mathbf{R}_x \left(\frac{\pi}{2} - \varphi_r^0 \right) \mathbf{R}_z \left(\frac{\pi}{2} + \lambda_r^0 \right) \\ &= \begin{bmatrix} -\sin \lambda_r^0 & \cos \lambda_r^0 & 0 \\ -\sin \varphi_r^0 \cos \lambda_r^0 & -\sin \varphi_r^0 \sin \lambda_r^0 & +\cos \varphi_r^0 \\ +\cos \varphi_r^0 \cos \lambda_r^0 & +\cos \varphi_r^0 \sin \lambda_r^0 & +\sin \varphi_r^0 \end{bmatrix} . \end{aligned} \quad (21.31)$$

Here $\mathbf{r}_{r,0} = [x_r^0, y_r^0, z_r^0]^\top$ denotes the approximated ECEF position of point r and $[\varphi_r^0, \lambda_r^0, h_r^0]$ its corresponding ellipsoidal coordinates. When working with ENU coordinates, the LOS vectors should be changed accordingly

$$\mathbf{e}_{r,1}^s = \mathbf{R}_x \left(\frac{\pi}{2} - \varphi_r^0 \right) \mathbf{R}_z \left(\frac{\pi}{2} + \lambda_r^0 \right) \mathbf{e}_r^s , \quad (21.32)$$

with $\mathbf{e}_{r,1}^s$ the LOS vector defined in the local ENU frame.

21.2.2 Linearized Observation Equations

Based on the expression for the linearized receiver–satellite range in (21.17) the *observed-minus-computed* counterparts of the observation equations (21.1) and (21.2) can be given as follows, first for code

$$\begin{aligned} \Delta p_{r,j}^s(t) = & -[\mathbf{e}_r^s(t)]^\top \Delta \mathbf{r}_r(t) + \Delta T_r^s(t) \\ & + [c - \dot{\rho}_r^s(t)] \Delta dt_r(t) \\ & + c [\Delta d_{r,j}^s(t) + \Delta \Delta d_{r,j}^s(t)] \\ & - c [\Delta dt^s(t - \tau_r^s) - \Delta d_j^s(t - \tau_r^s)] \\ & + \mu_j^S \Delta I_r^s(t) + e_{r,j}^s(t) . \end{aligned} \quad (21.33)$$

For the carrier-phase it can be given as

$$\begin{aligned} \Delta \varphi_{r,j}^s(t) = & -[\mathbf{e}_r^s(t)]^\top \Delta \mathbf{r}_r(t) + \Delta T_r^s(t) \\ & + [c - \dot{\rho}_r^s(t)] \Delta dt_r(t) \\ & + c [\Delta \delta_{r,j}^s(t) + \Delta \Delta \delta_{r,j}^s(t)] \\ & - c [\Delta dt^s(t - \tau_r^s) - \Delta \delta_j^s(t - \tau_r^s)] \\ & - \mu_j^S \Delta I_r^s(t) + \lambda_j^S \Delta N_{r,j}^s + \varepsilon_{r,j}^s(t) . \end{aligned} \quad (21.34)$$

The approximate values for the parameters that are needed in the linearization of the GNSS observation equations can usually be set to zero, if the linearized model is solved using *Gauss–Newton iteration* (Chap. 22). In case of zero approximate values for the unknown parameters, we can omit the Δ symbol in their notation. In order to limit the amount of iterations, in case of high-precision applications nonzero approximate values are used for the receiver position (these are usually available from SPP preprocessing, Sect. 21.3.4).

In both linearized code and phase-observation equations, the coefficient for the receiver clock error is $c - \dot{\rho}_r^s(t)$, that is, the speed of light minus the range rate, where the range rate is due to the linearization of the receiver–satellite range. As shown in the previous subsection, this range rate depends on the speed of the receiver. We can compute it based on approximate values of the receiver’s position and velocity, which can be obtained from a SPP solution at a previous epoch. Maximum range rate values for a static receiver on Earth is about 700 m/s, but for a spaceborne receiver in low-Earth orbit this is much higher, about 8000 m/s (Chap. 32). On the other hand, these values of the range rate are still very small compared to the speed of light c . And if the linearized model is solved in an iterative manner, in practice the results will be the same after several iterations are made compared to the results based on the observation equations neglecting the range rate [21.18]. In the positioning models treated in this chapter, we will therefore not include the range rate in the coefficients for the receiver clock error in the design matrix.

21.3 Point Positioning Models

The simplest positioning strategy is the one in which (single-frequency) pseudorange (code) observations measured by one receiver are processed to solve its position, given the positions of the GNSS satellites and their clock errors as computed from either the broadcast navigation message or precise (IGS) products. Some-

times ionospheric corrections can also be computed using information in the navigation message, or by an externally provided model. This strategy is referred to as single point positioning (SPP), and its solution is often referred to as *navigation solution*. SPP can also be carried out based on dual- or multifrequency code ob-

servations, thereby eliminating the ionospheric delays. Point positioning based on (single- or multifrequency) code and phase data, as well as precise products for orbits and clocks, is known as *precise point positioning* (PPP). Figure 21.3 visualizes the concept of point positioning.

21.3.1 Computation of the Satellite Clocks and Hardware Code (Group) Delays

In case of point positioning the satellite clock offsets need to be known. In case broadcast ephemeris (SPP) are used they can be computed using a polynomial model of which the coefficients are transmitted in the navigation message (see also the Interface Control Documents or ICDs). For GPS, Galileo, BeiDou, QZSS, and IRNSS the broadcast satellite clock can be calculated as the following second-order polynomial [21.22]

$$dt_S^s(t) = a_0^s(t_{oc}^s) + a_1^s(t - t_{oc}^s) + a_2^s(t - t_{oc}^s)^2 + \Delta t_{rel}^s(t), \quad (21.35)$$

with $S \in \{G, E, C, J, I\}$. The coefficients of the polynomial are denoted as a_0^s , a_1^s , and a_2^s , representing the offset, drift, and aging of the clock, and t_{oc}^s denotes the reference time of the clock data, which is also broadcast in the navigation message. The last part of the satellite clock corrections, denoted as $\Delta t_{rel}^s(t)$, is a *relativistic* correction, because the satellite clock is moving

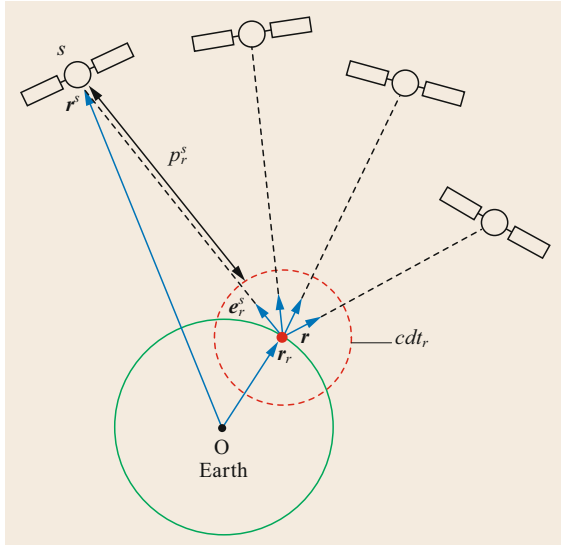


Fig. 21.3 Point positioning based on 4 GNSS satellites: r_r denotes the receiver position vector, r^s the satellite position vector, e_r^s the line-of-sight unit vector, cdt_r the receiver clock error, and p_r^s the code or pseudo-range observation

with respect to the receiver. This correction, which depends on the eccentric anomaly, has been discussed in Chap. 19. In the point-positioning algorithm the above clock offset polynomial needs to be evaluated at the time of transmission, that is, $t - \tau_r^s$.

GPS, Galileo, QZSS, and IRNSS

In case of GPS, Galileo, QZSS, and IRNSS, the broadcast satellite clock correction in fact corresponds to the *ionosphere-free combination* of dual-frequency code observations, see also Chap. 20 for this linear combination

$$dt_S^s(t) = dt_{IF}^s(t) = dt^s(t) - d_{IF}^s(t). \quad (21.36)$$

Here $d_{IF}^s(t)$ denotes the ionosphere-free combination of satellite hardware biases and is defined as

$$\begin{aligned} d_{IF}^s(t) &= \frac{\mu_2^S}{\mu_2^S - \mu_1^S} d_1^s(t) - \frac{\mu_1^S}{\mu_2^S - \mu_1^S} d_2^s(t) \\ &= d_j^s(t) + \frac{\mu_j^S}{\mu_2^S - \mu_1^S} \underbrace{[d_1^s(t) - d_2^s(t)]}_{DCB_{12}^s(t)}, \end{aligned} \quad (21.37)$$

for $j = 1, 2$.

The ionospheric coefficients corresponding to the two frequencies involved in the combination are denoted as μ_1^S and μ_2^S (21.3), where $S \in \{G, E, J, I\}$. Note that the ionosphere-free coefficients can also be given as function of the frequencies, that is,

$$\begin{aligned} \frac{\mu_2^S}{\mu_2^S - \mu_1^S} &= \frac{(f_1^S)^2}{(f_1^S)^2 - (f_2^S)^2} \quad \text{and} \\ \frac{\mu_1^S}{\mu_2^S - \mu_1^S} &= \frac{(f_2^S)^2}{(f_1^S)^2 - (f_2^S)^2}. \end{aligned}$$

The frequency-difference between the satellite hardware biases in (21.37), that is, $d_1^s(t) - d_2^s(t)$, is also known as *differential code bias* (DCB) [21.6], or *inter-frequency bias* (IFB) [21.3]. While for GPS (and QZSS) the ionosphere-free satellite clock applies to the L1 and L2 frequencies, in case of Galileo the broadcast satellite clock corresponds to either the ionosphere-free combination of E1+E5a, or to the combination of E1+E5b. Depending on the Galileo service that is used (Chap. 9), it follows which navigation message type and which ionosphere-free clock is transmitted: E1+E5a in case of the freely accessible navigation message (F/NAV) in the open service and E1+E5b in case of the integrity navigation message (I/NAV) in the safety-of-life service [21.23]. In case of IRNSS the ionosphere-free clock refers to the S-band and L5 frequencies [21.24].

Single-frequency SPP users employing the broadcast satellite clock offset cannot directly use the ionosphere-free satellite clock, but they need to apply another correction, the so-called timing group delay (TGD) difference between the two frequencies, which is defined as [21.25]

$$T_{GD}^s(t) = -\frac{\mu_1^s}{\mu_2^s - \mu_1^s} \underbrace{[d_1^s(t) - d_2^s(t)]}_{DCB_{12}^s(t)} . \quad (21.38)$$

Remind that the ionospheric coefficients read $\mu_1^s = 1$ and $\mu_2^s = (\lambda_2^s/\lambda_1^s)^2$. Thus, the TGD is a scaled version of the DCB. Note that the Galileo ICD speaks of BGD (broadcast group delay) instead of TGD. Using the group delays, the combination of satellite clock and hardware delay for the first two frequencies can be reconstructed as

$$dt_S^s(t) - \frac{\mu_j^s}{\mu_1^s} T_{GD}^s(t) = dt^s(t) - d_j^s(t), \quad j = 1, 2 . \quad (21.39)$$

In case of GPS users employing the L1 C/A code instead of the P1 code another correction should be taken into account, which accounts for the difference in hardware biases between the P1 and C/A code

$$dt_G^s(t) - T_{GD}^s(t) + DCB_{1c}^s(t) = dt^s(t) - d_c^s(t) , \quad (21.40)$$

with $DCB_{1c}^s(t) = d_1^s(t) - d_c^s(t)$ the DCB between the P1 and C/A code, where $d_c^s(t)$ denotes the hardware bias of the C/A code. This P1-C/A DCB is typically of the order of 2 ns (60 cm) [21.26]. This correction is however not transmitted in the GPS *legacy* navigation message (i.e., the NAV message, modulated on the C/A code), but is transmitted in the modernized civilian NAV (CNAV) messages modulated on the L2C and L5 signals, together with additional group delay corrections that are referred to as *intersignal corrections* (ISCs) [21.27, 28].

BeiDou (BDS)

In case of BeiDou the broadcast satellite clock is *not* referring to an ionosphere-free combination, but to the single-frequency B3 signal [21.29]

$$dt_C^s(t) = dt^s(t) - d_3^s(t) . \quad (21.41)$$

Single-frequency SPP users that use either the B1 or B2 frequency cannot directly use this broadcast satellite clock offset, but need to apply a TGD, depending on the frequency they use

$$\begin{aligned} T_{GD1}^s(t) &= d_1^s(t) - d_3^s(t) , \\ T_{GD2}^s(t) &= d_2^s(t) - d_3^s(t) . \end{aligned} \quad (21.42)$$

The BeiDou navigation message provides both these TGDs. Single-frequency B1 and B2 SPP users should apply the respective TGD such that

$$\begin{aligned} dt_C^s(t) - T_{GD1}^s(t) &= dt^s(t) - d_1^s(t) , \\ dt_C^s(t) - T_{GD2}^s(t) &= dt^s(t) - d_2^s(t) . \end{aligned} \quad (21.43)$$

Dual- or multifrequency BeiDou users can compute their ionosphere-free clock offsets

$$dt_C^s(t) - \frac{\mu_2^c}{\mu_2^c - \mu_1^c} T_{GD1}^s(t) + \frac{\mu_1^c}{\mu_2^c - \mu_1^c} T_{GD2}^s(t) , \quad (21.44)$$

for B1+B2, and

$$dt_C^s(t) - \frac{\mu_3^c}{\mu_3^c - \mu_1^c} T_{GD1}^s(t) \quad (21.45)$$

for B1+B3.

GLONASS

In case of GLONASS, the satellite clock offset is based on the L1 frequency and is calculated as follows from the navigation message [21.30]

$$dt_R^s(t) = a_0^s(t_{oc}^s) + a_1^s(t - t_{oc}^s) , \quad (21.46)$$

where it is noted that in the GLONASS ICD the reference time of the clock data is referred to as t_b^s . In contrast to GPS, Galileo, and BeiDou, the clock offset is for GLONASS computed as a first-order polynomial, where it is remarked that the broadcast parameters a_0^s and a_1^s not only account for the satellite clock offset and drift, but also for relativistic effects. A separate compensation for these effects is, therefore, not needed [21.31]. For GLONASS, the difference between satellite hardware delays on the L1 and L2 frequencies is broadcast as well [21.30]

$$\Delta\tau_n^s(t) = d_2^s(t) - d_1^s(t) . \quad (21.47)$$

The combination of satellite clock and hardware delay on GLONASS L1 is then obtained as follows

$$dt_R^s(t) = dt^s(t) - d_1^s(t) , \quad (21.48)$$

while its counterpart on the GLONASS L2 frequency is reconstructed as

$$dt_R^s(t) - \Delta\tau_n^s(t) = dt^s(t) - d_2^s(t) . \quad (21.49)$$

21.3.2 Some Remarks on the TGDs/DCBs

The DCB or TGD is initially determined by the satellite manufacturer before launch and can be revised by the GNSS's control segment [21.22]. For GPS, since 1999 JPL determines improved TGD values (as a byproduct of their ionospheric mapping) that are uploaded to the GPS satellites [21.32]. The size of these DCBs (for GPS/GLONASS) is less than 15 ns (4.5 m) [21.27], while for BeiDou less than 20 ns [21.33], and are normally very stable in time (at least over one day) [21.34, 35]. In case precise satellite clocks are used (PPP), these clocks are also based on the ionosphere-free combination [21.36] and hence single-frequency users should correct for the DCBs. Precise satellite DCBs are made available on a regular (daily) basis by the IGS, as part of their global ionospheric map (GIM) product [21.37]. As the constellation-mean value of the DCBs is set by IGS convention to zero [21.6], whereas the GPS broadcast group delays are referenced to an empirical absolute hardware bias [21.25], there is an offset between the DCBs based on the broadcast group delays and those published by the IGS. Galileo determines the group delay as part of the orbit determination and time synchronization (ODTS) process and, like GPS, applies a zero-mean condition to the whole constellation [21.33].

21.3.3 Computation/Estimation of the Atmospheric Errors

In case of single-frequency point positioning (SPP or PPP) the GNSS observations may be a-priori corrected for the ionospheric delays. These corrections can be calculated using the models broadcast in the navigation message, such as the Klobuchar model for GPS [21.38], or the NeQuick model for Galileo [21.39]. Otherwise, more precise ionospheric corrections can be extracted from GIM as produced by the IGS. In case of multifrequency point positioning, ionospheric corrections are not required, since the ionospheric delays can be estimated or eliminated from the data themselves. However, estimation (or elimination) weakens the model, which results in longer times for the solution to converge, so for fast multifrequency positioning (precise) ionospheric corrections are essential.

The tropospheric delays can usually be largely corrected for using models such as the *Saastamoinen* model [21.40]. If needed, residual tropospheric delays can be estimated in the model. For more detailed information on atmospheric models, we refer to Chap. 6.

21.3.4 Single-Constellation SPP Model

Direct (Analytical) SPP Solution

It is possible to compute a SPP solution based on four pseudoranges (in case of one constellation) in *analytical* (or closed) form, based on the nonlinear observation equations and without linearization/iteration and the need for approximate values. The closed-form solution can also be used to serve for fast computation of the approximate receiver position. We refer to the literature, where several approaches have been proposed [21.41–43].

Single-Frequency SPP Model

In general, with more than four satellites, say m_S satellites, tracked by a receiver, the SPP model of single-frequency (linearized) pseudorange (code) observation equations can be given for a single epoch as, making use of the calculated satellite orbits and clocks, satellite hardware delays and atmospheric delays

$$E \left(\begin{bmatrix} \Delta \tilde{p}_{r,j}^1(t) \\ \vdots \\ \Delta \tilde{p}_{r,j}^{m_S}(t) \end{bmatrix} \right) = \underbrace{\begin{bmatrix} -[e_r^1(t)]^\top & 1 \\ \vdots & \vdots \\ -[e_r^{m_S}(t)]^\top & 1 \end{bmatrix}}_{\mathbf{J}_0} \begin{bmatrix} \Delta \mathbf{r}_r(t) \\ c dt_{r,j}^S(t) \end{bmatrix}. \quad (21.50)$$

Here $E(\cdot)$ denotes the expectation operator and a *tilde* is used to denote the pseudorange observable that is corrected for orbit, clock, hardware biases, and atmospheric delays. A usual choice for SPP is $j = 1$, that is, the first frequency (in case of GPS: L1 C/A), but the model can be solved for other frequencies as well. Matrix \mathbf{J}_0 of dimension $m_S \times 4$ denotes the *Jacobian* (Chap. 22) or *design matrix*, which can be written in the following compact form

$$\mathbf{J}_0 = [\mathbf{G}_r^S(t), \mathbf{u}_{m_S}] , \quad (21.51)$$

Matrix $\mathbf{G}_r^S(t) = [-e_r^1(t), \dots, -e_r^{m_S}(t)]^\top$ of dimension $m_S \times 3$ contains the LOS unit direction vectors, while the m_S -vector of ones is defined as $\mathbf{u}_{m_S} = (1, \dots, 1)^\top$. Unknown parameters in the SPP model are the receiver position vector $\Delta \mathbf{r}_r(t)$ and the receiver clock error $dt_{r,j}(t)$. This estimable receiver clock is a combination

of the *true* receiver clock error plus the (frequency-dependent) receiver hardware delay, as both terms cannot be separated, that is,

$$dt_{r,j}^S(t) = dt_r(t) + d_{r,j}^S. \quad (21.52)$$

Here it is assumed that the receiver hardware bias is stable in time, such that it can be denoted without time stamp. It is noted that in case of GLONASS, the inter-channel hardware biases are so small compared to the noise of the pseudorange observations [21.31] that for the purpose of SPP they can be neglected. If the SPP model is solved in an iterative least-squares sense, the approximate receiver position can be taken zero (corresponding to the center of the Earth). After convergence, this yields the least-squares estimators $\hat{\mathbf{r}}_r(t)$ and $\hat{dt}_{r,j}^S(t)$.

The *redundancy* of an observation model is defined as the number of observables minus the number of estimable parameters. For the single-frequency SPP model it reads $m_S - (3 + 1) = m_S - 4$ (satellites). From this it follows that the model is solvable for $m_S \geq 4$.

Dual-Frequency SPP Model

In the presence of dual-frequency data, instead of modeling the ionospheric delays, a common procedure is to take the ionosphere-free combination to *eliminate* the ionospheric delays from the data and basically solve the single-frequency SPP model as in (21.50) but then based on the ionosphere-free observations and parameters. Table 21.2 shows the numerical values of the ionosphere-free coefficients for selected dual-frequency combinations of GPS (and QZSS), GLONASS, Galileo, BeiDou, and IRNSS observables.

Instead of working with ionosphere-free combinations, in a dual-frequency situation one could also work with a model that is, like the single-frequency model in (21.50), based on the *uncombined* observables, but

Table 21.2 Numerical values of ionosphere-free coefficients for several dual-frequency combinations of GPS (L#), GLONASS (G#), Galileo (E#), BeiDou (B#), and IRNSS (S+L5) observables

Signals	$\frac{\mu_2^S}{\mu_2^S - \mu_1^S} = \frac{(f_1^S)^2}{(f_1^S)^2 - (f_2^S)^2}$	$\frac{\mu_1^S}{\mu_2^S - \mu_1^S} = \frac{(f_2^S)^2}{(f_1^S)^2 - (f_2^S)^2}$
L1+L2	2.5457	1.5457
L1+L5	2.2606	1.2606
G1+G2	2.5312	1.5312
E1+E5a	2.2606	1.2606
E1+E5b	2.4220	1.4220
E1+E5	2.3380	1.3380
B1+B2	2.4872	1.4872
B1+B3	2.9437	1.9437
S+L5	1.2868	0.2868

which models the ionospheric delays as additional parameters. This model can, however, not be used directly in a least-squares adjustment, as its design matrix is *rank deficient*, which means that some of its columns are linear dependent. The rank deficiency can be overcome by application of the *S-system* or *datum* theory (see Chap. 22 for a general description of the theory of rank-defect least-squares). This means that instead of the original parameters as above, only certain *linear combinations* of parameters are estimable. However, the design matrix corresponding to these linear parameter combinations is of full rank.

Based on the (corrected) pseudorange observables, the full-rank dual-frequency SPP model can be given as

$$E \left(\begin{bmatrix} \Delta \tilde{\mathbf{p}}_{r,1}^S(t) \\ \Delta \tilde{\mathbf{p}}_{r,2}^S(t) \end{bmatrix} \right) = \begin{bmatrix} \mathbf{G}_r^S(t) & \mathbf{u}_{m_S} & \mu_1^S \mathbf{I}_{m_S} \\ \mathbf{G}_r^S(t) & \mathbf{u}_{m_S} & \mu_2^S \mathbf{I}_{m_S} \end{bmatrix} \begin{bmatrix} \Delta \mathbf{r}_r(t) \\ cd\tilde{\mathbf{t}}_r^S(t) \\ \tilde{\mathbf{I}}_r^S(t) \end{bmatrix}. \quad (21.53)$$

The estimable ionospheric delay parameters are stored as vector $\tilde{\mathbf{I}}_r^S(t) = [\tilde{I}_r^1(t), \dots, \tilde{I}_r^{m_S}(t)]^\top$. Inside the design matrix, \mathbf{I}_{m_S} denotes the identity matrix of dimension m_S , while vector \mathbf{u}_{m_S} and matrix $\mathbf{G}_r^S(t)$ are the same as in the single-frequency model.

Apart from the receiver position, the estimable receiver clock and ionospheric parameters have the following interpretation

$$\begin{aligned} d\tilde{\mathbf{t}}_r^S(t) &= dt_r(t) + d_{r,1}^S + \frac{\mu_1^S}{\mu_2^S - \mu_1^S} \text{DCB}_{r,12}^S, \\ \tilde{\mathbf{I}}_r^S(t) &= \mathbf{I}_r^S(t) - \frac{1}{\mu_2^S - \mu_1^S} c \text{DCB}_{r,12}^S. \end{aligned} \quad (21.54)$$

Here use is made of the following definition of a *receiver DCB* between the two frequencies

$$\text{DCB}_{r,12}^S = d_{r,1}^S - d_{r,2}^S. \quad (21.55)$$

Thus, the receiver DCB is *not* an estimable parameter in this model, but shows up as a *bias* in the interpretation of both the receiver clock and ionospheric parameters. This bias gets, however, eliminated when the observation equations are *reconstructed* from the estimable parameters, since for the first frequency it holds that

$$cd\tilde{\mathbf{t}}_r^S(t) + \mu_1^S \tilde{\mathbf{I}}_r^S(t) = c [dt_r(t) + d_{r,1}^S] + \mu_1^S \mathbf{I}_r^S(t). \quad (21.56)$$

For the second frequency it holds that

$$\begin{aligned} cd\tilde{\mathbf{t}}_r^S(t) + \mu_2^S \tilde{\mathbf{I}}_r^S(t) &= c [dt_r(t) + d_{r,1}^S] + \mu_2^S \mathbf{I}_r^S(t) \\ &\quad - c \text{DCB}_{r,12}^S = c [dt_r(t) + d_{r,2}^S] + \mu_2^S \mathbf{I}_r^S(t). \end{aligned} \quad (21.57)$$

We remark that the estimable receiver clock parameter in the dual-frequency SPP model can be rewritten as an *ionosphere-free* receiver clock

$$d\tilde{r}_r^S(t) = dt_r(t) + d_{r, \text{IF}}^S. \quad (21.58)$$

Here $d_{r, \text{IF}}^S$ denotes the ionosphere-free combination of the receiver code delays of the two frequencies (similar to that for the satellite code delays, see (21.37)).

The *redundancy* of the dual-frequency SPP model reads $2m_S - [3 + 1 + m_S] = m_S - 4$, which means that the model is solvable if $m_S \geq 4$, which is similar to the single-frequency SPP model.

21.3.5 Multiconstellation SPP Model

This subsection focuses on the combined multiconstellation SPP model. Like with the single-constellation SPP model, first one frequency per constellation is assumed, followed by two frequencies per constellation. These frequencies between the constellations may be identical, but can also be different.

SPP Model: One Frequency per Constellation

Suppose we have pseudorange data from *two* constellations, denoted as GNSS A tracking single-frequency data of m_A satellites and GNSS B tracking single-frequency data of m_B satellites. We can set up the following combined SPP model

$$\begin{aligned} E \left(\begin{bmatrix} \Delta \tilde{p}_{r,j}^A(t) \\ \Delta \tilde{p}_{r,j}^B(t) \end{bmatrix} \right) \\ = \begin{bmatrix} \mathbf{G}_r^A(t) & \mathbf{u}_{m_A} & \mathbf{0} \\ \mathbf{G}_r^B(t) & \mathbf{u}_{m_B} & \mathbf{u}_{m_B} \end{bmatrix} \begin{bmatrix} \Delta \mathbf{r}_r(t) \\ c dt_{r,j}^A(t) \\ c \text{ ISB}_{r,j}^{\text{AB}} \end{bmatrix}. \end{aligned} \quad (21.59)$$

The data of both constellations have the receiver coordinates in common, as well as the receiver clock, which is defined to be relative to the system time of constellation A. For the observations of constellation B an additional parameter shows up, which is the ISB

$$\text{ISB}_{r,j}^{\text{AB}} = [d_{r,j}^B - d_{r,j}^A] - t^{\text{AB}}. \quad (21.60)$$

Compared to the ISB definition given in (21.10), the interchannel terms are not present in the above equations, as they may be neglected for the purpose of SPP. Another difference is that in addition to the difference in receiver hardware delays of the signals of the two constellations, the estimable ISB parameter in case of SPP is biased by the time offset between the constellations, that is, t^{AB} (21.11). The reason is that it cannot be separated from the hardware delays difference and

therefore it is only estimable lumped to them. If this time offset is known (e.g., computed from the navigation message), it disappears from the interpretation of the ISB parameter. We furthermore remark that even if the frequencies of the signals of both constellations are identical (e.g., GPS L1 and Galileo E1), these ISBs do not cancel out [21.1, 44]. Instead of solving an ISB parameter, one may also introduce a receiver clock error corresponding to the second constellation in the SPP model. The following relating then holds between the receiver clocks and the ISB

$$\begin{aligned} dt_{r,j}^B(t) &= dt_{r,j}^A(t) + \text{ISB}_{r,j}^{\text{AB}} \\ &= dt_r(t) + d_{r,j}^B - t^{\text{AB}}. \end{aligned} \quad (21.61)$$

Based on this, another definition of the ISB can be given as

$$\text{ISB}_{r,j}^{\text{AB}} = dt_{r,j}^B(t) - dt_{r,j}^A(t), \quad (21.62)$$

that is, the difference of the estimable receiver clocks of the two constellations. Instead of parameterizing these constellation-specific receiver clocks, the ISB-parametrization is more advantageous in the event it is possible to *calibrate* the ISB. In that case, the ISB can be assumed known and the observations of constellation B are corrected for it, such that the multiconstellation SPP model becomes

$$E \left(\begin{bmatrix} \Delta \tilde{p}_{r,j}^A(t) \\ \Delta \tilde{p}_{r,j}^B(t)' \end{bmatrix} \right) = \begin{bmatrix} \mathbf{G}_r^A(t) & \mathbf{u}_{m_A} \\ \mathbf{G}_r^B(t) & \mathbf{u}_{m_B} \end{bmatrix} \begin{bmatrix} \Delta \mathbf{r}_r(t) \\ c dt_{r,j}^A(t) \end{bmatrix}, \quad (21.63)$$

with the ISB-corrected observables denoted as

$$\Delta \tilde{p}_{r,j}^B(t)' = \Delta \tilde{p}_{r,j}^B(t) - \mathbf{u}_{m_B} c \text{ ISB}_{r,j}^{\text{AB}}.$$

In the ISB-corrected model the observations of constellations A and B have the same parameters in common, and the dual-constellation model becomes equivalent to a *single-constellation* SPP model, but now based on $m_A + m_B$ satellites.

The *redundancy* of the combined SPP model based on one frequency per system reads, if the ISB is unknown, $m_A + m_B - 5$. This means that $m_A + m_B \geq 5$, and this can be satisfied by different combinations of satellites. In case of more than two constellations, model (21.59) is extendable with one combined ISB/time-offset parameter for each constellation that is added. The multiconstellation redundancy then reads in general $\sum_{i=1}^s m_i - (3 + s)$, where s denotes the number of constellations. In the case the ISB/time-offsets can be assumed known, the multiconstellation redundancy increases to $\sum_{i=1}^s m_i - 4$.

SPP Model: Two Frequencies per Constellation

The multiconstellation SPP model for two frequencies per constellation can be set up by modeling ionospheric delays as unknown parameters, as done in the single-constellation case in (21.53). In this case, the estimable parameters are, next to the receiver coordinates, an ionosphere-free clock for constellation A, as well as ionospheric delays per constellation that are biased by constellation-specific DCBs (21.54). In addition, an ISB parameter is parameterized, however in this case it is the *ionosphere-free* ISB parameter, defined as

$$\begin{aligned} \text{ISB}_{r, \text{IF}}^{\text{AB}} &= [d_{r, \text{IF}}^{\text{B}} - d_{r, \text{IF}}^{\text{A}}] - t^{\text{AB}} \\ &= \left[\frac{\mu_2^{\text{B}}}{\mu_2^{\text{B}} - \mu_1^{\text{B}}} d_{r,1}^{\text{B}} - \frac{\mu_1^{\text{B}}}{\mu_2^{\text{B}} - \mu_1^{\text{B}}} d_{r,2}^{\text{B}} \right] \\ &\quad - \left[\frac{\mu_2^{\text{A}}}{\mu_2^{\text{A}} - \mu_1^{\text{A}}} d_{r,1}^{\text{A}} - \frac{\mu_1^{\text{A}}}{\mu_2^{\text{A}} - \mu_1^{\text{A}}} d_{r,2}^{\text{A}} \right] \\ &\quad - t^{\text{AB}}. \end{aligned} \quad (21.64)$$

This ionosphere-free ISB corresponds to the definition given by [21.3, 7].

21.3.6 Precision and DOP

The impact of the receiver–satellite geometry (captured in matrix $\mathbf{G}_r^{\text{S}}(t)$ in case of a single constellation) on the precision of the receiver position obtained using SPP is usually described using the DOP concept [21.45]. The 4×4 cofactor matrix (variance matrix excluding the variance factor) of the receiver position and receiver clock of the single-constellation single-frequency model can analytically be given as [21.46]

$$(\mathbf{J}_0^{\top} \mathbf{J}_0)^{-1} = \begin{bmatrix} \mathbf{C}_{\hat{r}(t)}^{\text{S}} & \mathbf{C}_{\hat{r}(t)}^{\text{S}} \bar{\mathbf{e}}_r^{\text{S}}(t) \\ \bar{\mathbf{e}}_r^{\text{S}}(t)^{\top} \mathbf{C}_{\hat{r}(t)}^{\text{S}} & \frac{1}{m_{\text{S}}} + \bar{\mathbf{e}}_r^{\text{S}}(t)^{\top} \mathbf{C}_{\hat{r}(t)}^{\text{S}} \bar{\mathbf{e}}_r^{\text{S}}(t) \end{bmatrix}. \quad (21.65)$$

This cofactor matrix is based on the design matrix of the SPP model given as in (21.51) and where

$$\bar{\mathbf{e}}_r^{\text{S}}(t) = \frac{1}{m_{\text{S}}} \sum_{s=1}^{m_{\text{S}}} \mathbf{e}_r^{\text{S}}(t)$$

denotes the mean LOS vector over all satellites. The 3×3 cofactor matrix of the receiver position is given as

$$\mathbf{C}_{\hat{r}(t)}^{\text{S}} = \left(\sum_{s=1}^{m_{\text{S}}} [\mathbf{e}_r^{\text{S}}(t) - \bar{\mathbf{e}}_r^{\text{S}}(t)] [\mathbf{e}_r^{\text{S}}(t) - \bar{\mathbf{e}}_r^{\text{S}}(t)]^{\top} \right)^{-1}. \quad (21.66)$$

From this last expression, one can see that if the satellite LOS vectors differ a lot from each other and from their mean, that is, when $\mathbf{e}_r^{\text{S}}(t) - \bar{\mathbf{e}}_r^{\text{S}}(t)$ is large, this is favorable for the precision with which the receiver position can be determined.

DOP values can now be computed based on the diagonal elements of the cofactor matrix of the receiver position. If $\mathbf{r}(t) = [E, N, U]^{\top}$ denotes the position in a local east–north–up frame, then the following DOPs can be calculated

$$\begin{aligned} \text{GDOP} &= \sqrt{c_{\hat{E}}^2 + c_{\hat{N}}^2 + c_{\hat{U}}^2 + c_{\text{cd}\hat{r},j}^{\text{S}}}, \\ \text{PDOP} &= \sqrt{c_{\hat{E}}^2 + c_{\hat{N}}^2 + c_{\hat{U}}^2}, \\ \text{HDOP} &= \sqrt{c_{\hat{E}}^2 + c_{\hat{N}}^2}, \\ \text{VDOP} &= \sqrt{c_{\hat{U}}^2}. \end{aligned} \quad (21.67)$$

Here

$$c_{\hat{E}}^2, c_{\hat{N}}^2, \quad \text{and} \quad c_{\hat{U}}^2$$

denote the diagonal elements of $\mathbf{C}_{\hat{r}}(t)$ for the position, whereas

$$c_{\text{cd}\hat{r},j}^{\text{S}}$$

denotes the diagonal element of (21.65) for the receiver clock. GDOP stands for *geometric* dilution of precision, PDOP for *position* dilution of precision, HDOP for *horizontal* dilution of precision, and VDOP for *vertical* dilution of precision. Unfavorable geometries, however, may lead to poor receiver precision and in some unfortunate cases even the position cannot be determined (geometry singularities). For example, when the end points of the LOS vectors describe a plane (all satellites lie on the surface of a cone) [21.47]. In that case the DOPs are infinitely large.

In case of *multiconstellation* positioning, it is also possible to calculate DOP values. If we assume two constellations, A and B, then these DOPs are computed based on the dual-constellation model (21.59), for which the receiver position cofactor matrix follows as

$$\mathbf{C}_{\hat{r}(t)} = \left[\left(\mathbf{C}_{\hat{r}(t)}^{\text{A}} \right)^{-1} + \left(\mathbf{C}_{\hat{r}(t)}^{\text{B}} \right)^{-1} \right]^{-1}, \quad (21.68)$$

that is, the inverse of a *sum* of the inverse cofactor matrices according to (21.66) that correspond to each individual constellation. From this expression, it easily follows that the dual-constellation DOPs are smaller (or in the worst case: equal) than their single-constellation

counterparts. In case both ISB and time offset are a priori *known*, the receiver position cofactor matrix follows from model (21.63) as

$$\mathbf{C}_r(t) = \left(\sum_{s=1}^{m_A+m_B} [\mathbf{e}_r^s(t) - \bar{\mathbf{e}}_r(t)] [\mathbf{e}_r^s(t) - \bar{\mathbf{e}}_r(t)]^\top \right)^{-1}, \quad (21.69)$$

with the mean LOS vector taken over *both* constellations, that is,

$$\bar{\mathbf{e}}_r(t) = \frac{1}{m_A + m_B} \sum_{s=1}^{m_A+m_B} \mathbf{e}_r^s(t).$$

The DOP values based on this model are even smaller than those based on (21.68).

As an example, Fig. 21.4 depicts PDOP values for a certain receiver–satellite geometry of four GPS and four Galileo satellites. Besides constellation-specific PDOPs, the PDOPs are shown for the combined GPS+Galileo model assuming a receiver clock and ISB parameter (similar to assuming a receiver clock per constellation), as well as PDOPs when only one receiver clock parameter is assumed for both GPS and Galileo (ISBs known). The PDOPs of the combined model with one receiver clock are the smallest, as this model is the strongest, although in Fig. 21.4 the PDOPs of this model tend to become equal to the PDOPs of the model with two receiver clocks, but this is due to the actual geometry in this example.

21.3.7 PPP Model

If, in addition to the pseudoranges, also the carrier-phase observations are employed for point positioning, in combination with the use of precise (IGS) products, we speak of PPP [21.48, 49]. For the full details and intricacies of the PPP technique, we refer to Chap. 25; here we will restrict ourselves to an overview of the single-frequency and multifrequency PPP models based on one constellation, as to compare to their SPP counterparts. Figure 21.5 visualizes the procedure for PPP, where in the first step a reference network (e.g., the IGS network) determines (satellite-dependent) parameters that are in a second step transmitted to the users. In the third step a user applies this correction information which enables PPP.

PPP users employing the precise (IGS) products should be aware that the precise satellite clocks are based on the *ionosphere-free* combination, similar to the broadcast satellite clock for most constellations. In case of single-frequency PPP, therefore, corrections for the satellite DCBs are required, but these are also provided by the IGS or its analysis centers. In addition, as single-frequency PPP users cannot form the ionosphere-free dual-frequency combination, ionospheric corrections are essential and these can be obtained from the IGS as well, in the GIM format.

Concerning the tropospheric delays, in addition the a-priori model corrections, for precise positioning applications it may be necessary to parameterize (residual) tropospheric delays. A common procedure is to

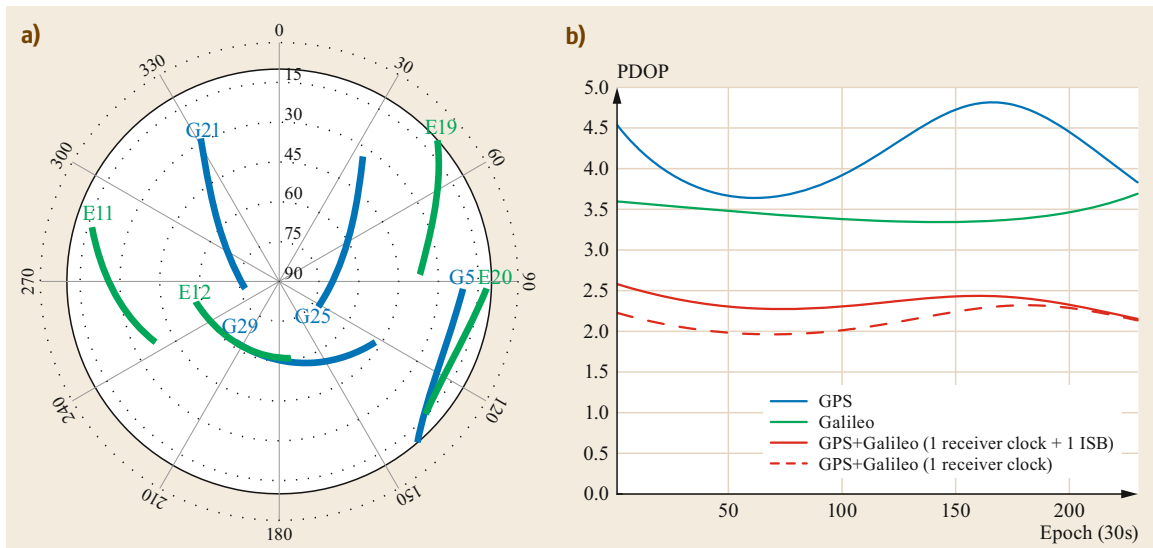


Fig. 21.4a,b Skyplot (a) and PDOP values (b), based on 4 Galileo-IOV and 4 GPS satellites, above 10 deg cut-off elevation in Perth, Australia, for 04:05–6:00 GPST on 20 March 2013. For the combined GPS+Galileo case, distinction is made between the PDOPs based on model (21.59) and model (21.63)

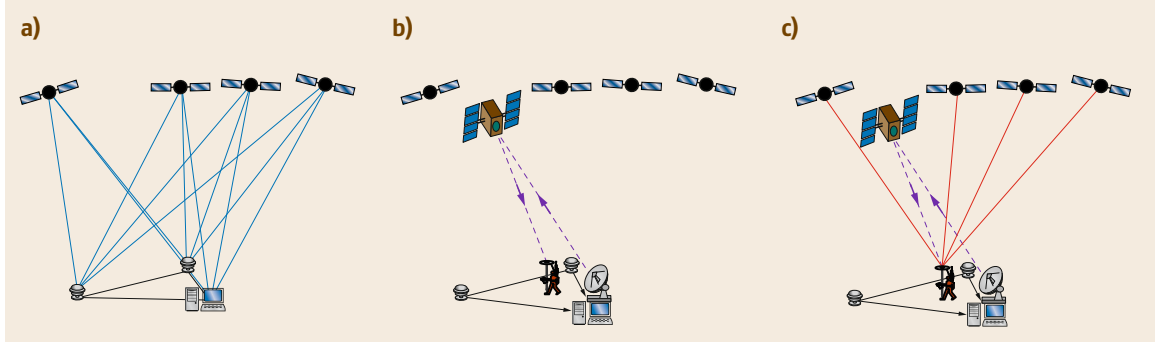


Fig. 21.5a–c Visualization of the PPP(-RTK) concept: **(a)** CORS (global or regional) network determines GNSS parameters; **(b)** satellite-dependent parameters are uploaded by the network and downloaded by a user; **(c)** the user applies the corrections to his data, enabling single-receiver precise positioning

map the residual tropospheric delays to local zenith, that is,

$$T_r^s(t) = T_{r,0}^s(t) + m_r^s(t)T_r^z(t),$$

with $T_{r,0}^s(t)$ the a-priori tropospheric correction, $m_r^s(t)$ the mapping function, and T_r^z the zenith tropospheric delay (ZTD). An example of an accurate tropospheric mapping function is *Niell's* mapping function [21.50].

GLONASS PPP (and also RTK) requires a-priori correction of the receiver- and frequency-dependent interchannel or interfrequency biases, at least for the phase data, that is, $\Delta\delta_{r,j}^s$. In [21.51], a table is presented with interchannel corrections for GLONASS receivers of nine different manufacturers. Here we assume that the data are a-priori corrected for these biases, such that GLONASS data can be processed using the general models we present here.

Due to a lack of space, the PPP models discussed here are restricted to one GNSS constellation only. PPP models for multiple constellations can however be developed along similar lines as the multiconstellation SPP model in Sect. 21.3.5. For notational convenience, from now on the system identifier S will be omitted in a single-constellation case.

Single-Frequency PPP Model

In case of single-frequency PPP it is assumed that identical offsets for the satellite clock and hardware bias apply to both code and phase observations, as well as a-priori corrections for tropospheric and ionospheric delays. Provided that the satellite clock offsets are based on the ionosphere-free combination (as is the case with precise IGS products), the offset terms for code and phase then can be given as

$$o_{p_{r,j}}^s(t) = c \left[dt_{IF}^s(t - \tau_r^s) + \frac{\mu_j}{\mu_2 - \mu_1} \text{DCB}_{12}^s \right] - T_{r,0}^s(t) - \mu_j T_r^s(t),$$

$$o_{\varphi_{r,j}}^s(t) = c \left[dt_{IF}^s(t - \tau_r^s) + \frac{\mu_j}{\mu_2 - \mu_1} \text{DCB}_{12}^s \right] - T_{r,0}^s(t) + \mu_j T_r^s(t). \quad (21.70)$$

Like the receiver hardware biases, it is assumed that the satellite hardware biases are stable such that they can be kept as time constants in the model. The DCBs (difference of satellite hardware biases) are needed to convert the ionosphere-free satellite clocks to the clocks plus hardware bias for the required frequency. It is remarked that in case of GPS the code corresponds to the C/A code ($j = 1$), we also need to subtract the P1-C/A DCB (i. e., DCB_{1c}) from the code correction. Applying the above offsets to the single-frequency code and phase observations yields the (linearized) full-rank single-frequency PPP model as given in Table 21.3 (top row).

The corrected (observed-minus-computed) code and phase observables then read $\Delta\tilde{p}_{r,j}(t) = \Delta p_{r,j}(t) + o_{p_{r,j}}^s(t)$ and $\Delta\tilde{\varphi}_{r,j}(t) = \Delta\varphi_{r,j}(t) + o_{\varphi_{r,j}}^s(t)$, respectively. Precise satellite orbits are used to calculate the LOS vectors in geometry matrix $\mathbf{G}_r(t)$ and to compute approximate values for the receiver–satellite ranges, used in the linearization. In this PPP model, the (residual) ZTD parameter is combined with the position parameters in the four-dimensional vector $\mathbf{x}_r(t)$, defined as

$$\mathbf{x}_r(t) = [\Delta\mathbf{r}_r(t)^\top, T_r^z(t)]^\top. \quad (21.71)$$

In addition, the LOS vectors and tropospheric mapping coefficients are stored in the $m \times 4$ matrix $\mathbf{G}_r(t) = [\mathbf{g}_r^1(t), \dots, \mathbf{g}_r^m(t)]^\top$, with the 4×1 geometry vector $\mathbf{g}_r^s(t)$ now consisting of the LOS vector, *plus* the tropospheric mapping function coefficient, which is defined as

$$\mathbf{g}_r^s(t) = \begin{bmatrix} -\mathbf{e}_r^s(t) \\ m_r^s(t) \end{bmatrix}. \quad (21.72)$$

Besides the receiver position, unknown parameter for both code and phase is the (biased) receiver clock $\tilde{d}_r(t)$,

Table 21.3 Full-rank undifferenced PPP models and their estimable parameters. Note: SF = single-frequency; DF = dual-frequency

Model	Notation and interpretation of estimable parameters
SF PPP ionosphere-corrected	$E \left(\begin{bmatrix} \Delta \tilde{\mathbf{p}}_{r,j}(t) \\ \Delta \tilde{\varphi}_{r,j}(t) \end{bmatrix} \right) = \begin{bmatrix} \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \mathbf{0} \\ \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{u}_m & \lambda_j \mathbf{C}_m \end{bmatrix} \begin{bmatrix} \mathbf{x}_r(t) \\ cd\tilde{I}_r(t) \\ c\tilde{\delta}_{r,j} \\ \tilde{\mathbf{N}}_{r,j} \end{bmatrix}$ <p>The following parameters are estimable ($j \in \{1, 2\}$):</p> $d\tilde{I}_r(t) = dI_r(t) + d_{r,j}$ $\tilde{\delta}_{r,j} = \delta_{r,j} - d_{r,j} + \frac{\lambda_j}{c} \left(N_{r,j}^p + \frac{c}{\lambda_j} [\delta_j^p - d_j^p] \right)$ $\tilde{\mathbf{N}}_{r,j}^s = \left(N_{r,j}^s + \frac{c}{\lambda_j} [\delta_j^s - d_j^s] \right) - \left(N_{r,j}^p + \frac{c}{\lambda_j} [\delta_j^p - d_j^p] \right)$
SF PPP ionosphere-float	$E \left(\begin{bmatrix} \Delta \tilde{\mathbf{p}}_{r,j}(t_1) \\ \Delta \tilde{\mathbf{p}}_{r,j}(t_2) \\ \Delta \tilde{\varphi}_{r,j}(t_1) \\ \Delta \tilde{\varphi}_{r,j}(t_2) \end{bmatrix} \right) = \begin{bmatrix} \mathbf{G}_r(t_1) & \mathbf{0} & \mathbf{u}_m & \mathbf{0} & \mu_j \mathbf{I}_m & \mathbf{0} & \mathbf{0} \\ \mathbf{G}_r(t_1) & \mathbf{0} & \mathbf{u}_m & \mathbf{0} & -\mu_j \mathbf{I}_m & \mathbf{0} & \lambda_j \mathbf{C}_m \\ \mathbf{0} & \mathbf{G}_r(t_2) & \mathbf{0} & \mathbf{u}_m & \mathbf{0} & \mu_j \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_r(t_2) & \mathbf{0} & \mathbf{u}_m & \mathbf{0} & -\mu_j \mathbf{I}_m & \lambda_j \mathbf{C}_m \end{bmatrix} \begin{bmatrix} \mathbf{x}_r(t_1) \\ \mathbf{x}_r(t_2) \\ cd\tilde{I}_r(t_1) \\ cd\tilde{I}_r(t_2) \\ \tilde{\mathbf{I}}_r(t_1) \\ \tilde{\mathbf{I}}_r(t_2) \\ \tilde{\mathbf{N}}_{r,j} \end{bmatrix}$ <p>The following parameters are estimable ($j \in \{1, 2\}$ and $i = 1, 2$):</p> $d\tilde{I}_r(t_i) = dI_r(t_i) + \frac{1}{2} \left[d_{r,j} + \delta_{r,j} + \frac{\lambda_j}{c} \left(N_{r,j}^p + \frac{c}{\lambda_j} [\delta_j^p - d_j^p] \right) \right]$ $\tilde{I}_r^s(t_i) = I_r^s(t_i) + \frac{1}{2\mu_j} c \left[d_{r,j} - \delta_{r,j} - \frac{\lambda_j}{c} \left(N_{r,j}^p + \frac{c}{\lambda_j} [\delta_j^p - d_j^p] \right) \right]$ $\tilde{\mathbf{N}}_{r,j}^s = \left(N_{r,j}^s + \frac{c}{\lambda_j} [\delta_j^s - d_j^s] \right) - \left(N_{r,j}^p + \frac{c}{\lambda_j} [\delta_j^p - d_j^p] \right)$
DF PPP ionosphere-float	$E \left(\begin{bmatrix} \Delta \tilde{\mathbf{p}}_{r,1}(t) \\ \Delta \tilde{\mathbf{p}}_{r,2}(t) \\ \Delta \tilde{\varphi}_{r,1}(t) \\ \Delta \tilde{\varphi}_{r,2}(t) \end{bmatrix} \right) = \begin{bmatrix} \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \mathbf{0} & \mu_1 \mathbf{I}_m & \mathbf{0} & \mathbf{0} \\ \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \mathbf{0} & \mu_2 \mathbf{I}_m & \mathbf{0} & \mathbf{0} \\ \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{u}_m & \mathbf{0} & -\mu_1 \mathbf{I}_m & \lambda_1 \mathbf{C}_m & \mathbf{0} \\ \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \mathbf{u}_m & -\mu_2 \mathbf{I}_m & \mathbf{0} & \lambda_2 \mathbf{C}_m \end{bmatrix} \begin{bmatrix} \mathbf{x}_r(t) \\ cd\tilde{I}_r(t) \\ c\tilde{\delta}_{r,1} \\ c\tilde{\delta}_{r,2} \\ \tilde{\mathbf{I}}_r(t) \\ \tilde{\mathbf{N}}_{r,1} \\ \tilde{\mathbf{N}}_{r,2} \end{bmatrix}$ <p>The following parameters are estimable ($j = 1, 2$):</p> $d\tilde{I}_r(t) = dI_r(t) + d_{r,1F}$ $\tilde{\delta}_{r,j} = \delta_{r,j} - \left[d_{r,1F} + \frac{\mu_j}{\mu_2 - \mu_1} \text{DCB}_{r,12} \right] + \frac{\lambda_j}{c} \left(N_{r,j}^p + \frac{c}{\lambda_j} [\delta_j^p - d_{1F}^p - \frac{\mu_j}{\mu_2 - \mu_1} \text{DCB}_{12}^p] \right)$ $\tilde{I}_r^s(t) = I_r^s(t) - \frac{1}{\mu_2 - \mu_1} c [\text{DCB}_{12}^s + \text{DCB}_{r,12}]$ $\tilde{\mathbf{N}}_{r,j}^s = \left(N_{r,j}^s + \frac{c}{\lambda_j} [\delta_j^s - d_{1F}^s - \frac{\mu_j}{\mu_2 - \mu_1} \text{DCB}_{12}^s] \right) - \left(N_{r,j}^p + \frac{c}{\lambda_j} [\delta_j^p - d_{1F}^p - \frac{\mu_j}{\mu_2 - \mu_1} \text{DCB}_{12}^p] \right)$

of which its interpretation is identical to that of single-frequency SPP.

The phase observables in the PPP model introduce their own specific parameters, which are a receiver-phase bias and phase ambiguity parameters. Unfortunately these parameters cannot be estimated independently, as their columns are linear dependent. To overcome this rank deficiency (of size 1), a choice is to estimate the *between-satellite differences* of the (biased) ambiguity parameters, instead of their undif-

ferenced counterparts, that is, the estimable ambiguity parameter is $\tilde{\mathbf{N}}_{r,j}^s$ in Table 21.3, for $s = 1, \dots, m$ and where $s \neq p$. With this reparameterization, there is thus one ambiguity parameter less estimable, since we form between-satellite differences with respect to the p th satellite. This arbitrarily chosen satellite is referred to as *pivot satellite*. Thus, the $(m-1)$ -vector of estimable ambiguities reads

$$\tilde{\mathbf{N}}_{r,j} = [\tilde{\mathbf{N}}_{r,j}^1, \dots, \tilde{\mathbf{N}}_{r,j}^{p-1}, \tilde{\mathbf{N}}_{r,j}^{p+1}, \dots, \tilde{\mathbf{N}}_{r,j}^m]^\top. \quad (21.73)$$

In the PPP design matrix in Table 21.3 this vector is multiplied by the $m \times (m-1)$ matrix \mathbf{C}_m , which is defined as

$$\mathbf{C}_m = \begin{bmatrix} \mathbf{I}_{p-1} & \mathbf{0} \\ \mathbf{0} & \begin{pmatrix} \mathbf{0}_{1 \times (m-p)} \\ \mathbf{I}_{m-p} \end{pmatrix} \end{bmatrix}. \quad (21.74)$$

This matrix can be regarded as the identity matrix of dimension m having its p -th column removed. A consequence of parameterizing between-satellite ambiguity differences for undifferenced observables is that it should be somewhere *compensated* by the other estimable parameters. In this case the estimable receiver-phase bias parameter, that is, $\delta_{r,j}$ in Table 21.3, gets biased by the ambiguity plus hardware biases corresponding to the pivot satellite.

Although the ambiguities $N_{r,j}^s$ have the property of being integer, in the PPP model they are *not* estimable as such, because of the lumping of the satellite-phase and code hardware delays to them (Table 21.3). The consequence is that for every epoch of data the number of phase observations equals the number of phase parameters they introduce, which means that the phase data do *not* contribute to the estimation of the receiver position, which is thus fully governed by the (less precise) code observables. However, this situation changes in *multiepoch* mode; in that case the phase data start to contribute to the solution of the receiver position (as their estimable ambiguity parameters are time constant). The longer this time span, the more it is governed by the phase data; after a certain time the position precision will have converged to a certain level. The *redundancy* of this multiepoch, single-frequency PPP model equals $m-5+(k-1)(2m-5)$, with k denoting the number of epochs. Note that for a single epoch ($k=1$) this redundancy reduces to $m-5$, which is 1 less than the redundancy of its SPP counterpart. This is due to the parameterization of the ZTD in the PPP model.

In the absence of ionospheric corrections, single-frequency PPP is still possible, thereby making use of the opposite sign of the ionospheric delays for code and phase. In this case the model is, however, not solvable based on a single epoch of data, as there are too many unknown parameters. Based on *two* epochs, however, the full-rank single-frequency PPP model is presented in Table 21.3 (second row). The estimable ambiguity parameters of this *ionosphere-float* model (as ionospheric delays are unknown parameters) have the same interpretation as in the single-frequency ionosphere-corrected PPP model, but the estimable receiver clock has a completely different interpretation; in the

ionosphere-float case it is biased by the receiver bias, as well as the pivot satellite ambiguity and hardware biases (there is no estimable receiver bias parameter). The estimable ionospheric parameter is biased by hardware delays and (pivot satellite) ambiguities as well. Based on a minimum of two epochs, the *redundancy* of this single-frequency, ionosphere-float PPP model equals $m-9$, requiring this model a large number of 9 satellites to be solvable. This redundancy is, however, based on the parameterization of different receiver positions for both epochs (i.e., a kinematic solution). If the receiver can be assumed *static* and the ZTD is assumed time constant, the model becomes stronger since $\mathbf{x}_r(t_1) = \mathbf{x}_r(t_2)$, increasing the redundancy with 4 to $m-5$, which is identical to the redundancy of the single-epoch, single-frequency, ionosphere-corrected PPP model.

Subject to the availability of precise GNSS orbits, clocks, DCBs, and ionosphere products, the *positioning accuracy* of single-frequency PPP is typically at decimeter level after a few minutes of convergence time [21.52].

Dual-Frequency PPP Model

In the dual-frequency PPP case, the ionospheric delays are assumed as unknown parameters and therefore no ionospheric corrections are incorporated. In that case DCBs are not needed as well. The correction terms for code and phase are in that case equivalent and read

$$o_{p,r,IF}^s(t) = o_{\varphi,r,IF}^s(t) = cdt_{IF}^s(t - \tau_r^s) - T_{r,0}^s(t). \quad (21.75)$$

Like in the single-frequency GPS case, if the first frequency corresponds to the C/A code ($j=1$), we need to subtract the P1-C/A DCB (i.e., DCB_{1c}) from the code correction. The linearized full-rank model for the dual-frequency PPP case is given in Table 21.3 (bottom row).

The estimability and interpretation of the code-related parameters (i.e., receiver position and clock) in this dual-frequency PPP model is exactly the same as in the dual-frequency SPP model (21.54) and (21.58). As a consequence of the absence of the satellite DCB corrections in the dual-frequency PPP model, satellite DCBs get lumped to the estimable ionospheric parameters and hence its interpretation becomes a combination of the true ionospheric delay plus satellite and receiver DCBs (Table 21.3). The ambiguity parameters also get biased by the satellite DCBs.

The phase-ambiguity parameters are in the dual-frequency PPP case estimable as between-satellite differenced parameters (relative to pivot satellite p), simi-

lar as in the single-frequency PPP model; however the interpretation between the dual- and single-frequency cases differs (Table 21.3). In both cases, the ambiguities are not integer estimable.

The *redundancy* of the dual-frequency PPP model equals $m - 5 + (k - 1)(3m - 5)$, which for a single epoch

($k = 1$) reduces to $m - 5$, similar to the previously discussed single-frequency PPP models. The *positioning accuracy* of dual-frequency (GPS) PPP can reach centimeter level, however only after a convergence time (based on the constant ambiguity terms) of typically more than 30 min [21.52].

21.4 Relative Positioning Models

In this section, we discuss the differential or relative GNSS positioning models, in which observations of more than one receiver are combined such that errors that are common between the receivers can be eliminated or reduced. Another important advantage is that in a relative measurement setup the carrier-phase ambiguities can be estimated to integer values, thereby greatly improving the positioning accuracy. The models in this section are restricted to the single-epoch case. For all relative positioning models it is assumed that in case of FDMA constellations interchannel bias corrections are a-priori corrected and do not show up in the observation equations. As with the models for PPP, the discussion of the relative positioning models in this chapter will be restricted to a single GNSS constellation only.

21.4.1 Principle of DGNSS and (PPP)-RTK

The typical accuracy of (single-constellation) SPP is in the order of 10 m. This accuracy is basically due to the uncertainty in the orbits, satellite clocks, and atmospheric delays. Despite the developments of PPP during the past decade, the technique of differential GNSS (DGNSS) has already been applied for decades to improve this positioning accuracy through eliminating or significantly removing errors that are common for receivers simultaneously tracking data of the same GNSS satellites. More details about DGNSS and services based on this concept can be found in Chap. 26. Here, we will briefly review the principle of DGNSS, following the discussion in [21.12].

Satellite Clock Evaluation

Due to the difference in travel time between two or more receivers in a relative positioning setup, the satellite clock and satellite hardware delay are evaluated at (slightly) different times of transmission, that is, $t - \tau_1^s$ for the reference receiver, denoted using subscript 1, and $t - \tau_r^s$ for the other (rover) receiver, denoted using subscript r (Fig. 21.6). This difference in travel

time, that is, $|\tau_r^s - \tau_1^s|$, is at most 19 ms (for one receiver experiencing the satellite in zenith and the other receiver experiencing the same satellite at zero degrees elevation) [21.18]. This means that the difference in transmission time, that is, $|(t - \tau_1^s) - (t - \tau_r^s)|$, is at most 19 ms as well. Assuming that the satellite clock drifts with a rate of at most 10^{-11} s/s (Chap. 5), this means that after 19 ms the satellite clock (in terms of distance) has changed with $c \cdot 10^{-11} \cdot 19 \cdot 10^{-3} \approx 0.06$ mm, which is negligible compared to the precision of the phase and code observations. Thus, we may safely assume that for the purpose of evaluation of the satellite clock at the

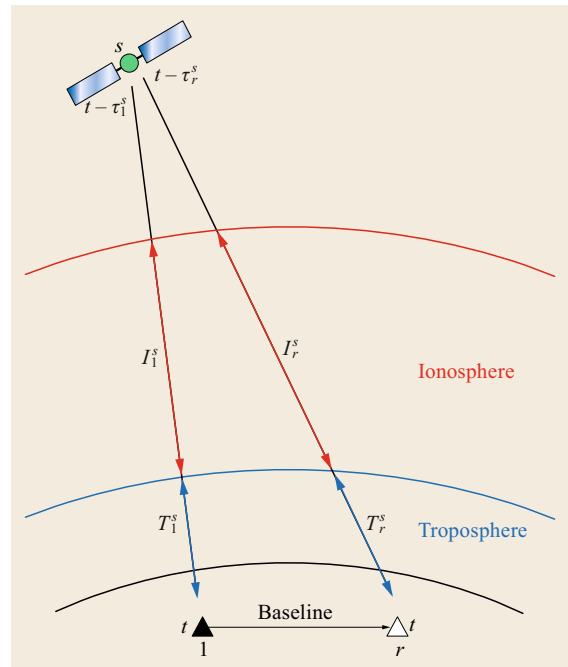


Fig. 21.6 Relative GNSS positioning: two receivers (reference 1 and rover r) quasi-simultaneously receive signals from satellite s at time t . On their way, they propagate through the atmosphere (ionosphere and atmosphere), affecting their travel times (denoted as τ_1^s and τ_r^s)

time of transmission

$$dt^s(t - \tau_1^s) = dt^s(t - \tau_r^s) \doteq dt^s(t - \tau^s). \quad (21.76)$$

Code-Dominated DGNSS Positioning

Recall the (nonlinear) code observation equation (21.1). Its expectation can be given as

$$\begin{aligned} E(p_{r,j}^s(t)) &= \rho_r^s(t, t - \tau_r^s) + T_r^s(t) + \mu_j I_r^s(t) \\ &\quad + c[dt_r(t) + d_{r,j}] \\ &\quad - c[dt^s(t - \tau_r^s) - d_j^s]. \end{aligned} \quad (21.77)$$

Now assume a reference receiver $r = 1$, which is stationed at a *known* location. Based on this known position together with the known position of the satellite (computed from the ephemeris), the range $\rho_1^s(t, t - \tau_1^s)$ can be computed. Subtracting the observed pseudorange from this computed range yields the *pseudorange corrections* (PRC) for satellite s

$$\text{PRC}_{p,j}^s(t) = \rho_1^s(t, t - \tau_1^s) - p_{1,j}^s(t). \quad (21.78)$$

Applying these pseudorange corrections to the pseudoranges of a user r yields, making use of (21.76)

$$\begin{aligned} E(\tilde{p}_{r,j}^s(t)) &= \rho_r^s(t, t - \tau_r^s) + T_{1r}^s(t) + \mu_j I_{1r}^s(t) \\ &\quad + c[dt_{1r}(t) + d_{1r,j}]. \end{aligned} \quad (21.79)$$

Here the corrected pseudorange reads

$$\tilde{p}_{r,j}^s(t) = p_{r,j}^s(t) + \text{PRC}_{p,j}^s(t).$$

Furthermore, *between-receiver differenced* unknown parameters are denoted as $(\cdot)_{1r} = (\cdot)_r - (\cdot)_1$. Thus, besides the elimination of satellite clocks and hardware delays, users employing the pseudorange corrections may significantly reduce the errors due to tropospheric and ionospheric delays. For sufficiently *short* distances, these differential atmospheric errors are so small (due to the spatial correlation of the atmosphere), compared to the measurement precision of the code data, that they may be neglected and disappear as unknown parameters. The combined differential receiver clock and hardware delay, that is, $c[dt_{1r}(t) + d_{1r,j}(t)]$, can be regarded as a receiver clock error to be solved by the user using a model which has the same structure as the SPP model discussed in the previous section.

Next to the pseudorange corrections, in practice usually also so-called *range-rate corrections* (RRC) are determined and transmitted (in real-time) to users, as to account for the difference between the time of determination of the corrections at the reference station (t_0) and the time the corrections are applied by the users (t)

$$\text{PRC}_{p,j}^s(t) = \text{PRC}_{p,j}^s(t_0) + (t - t_0)\text{RRC}_{p,j}^s(t_0), \quad (21.80)$$

where $t - t_0$ is referred to as *latency*. It will be clear that the accuracy of the pseudorange corrections improves for smaller latencies.

Using the above concept of DGNSS, the positioning accuracy can be improved to 1–2 m. The accuracy can be improved further (to submeter level) by using *carrier-phase smoothing* (Chap. 20), but its performance is limited due to local receiver bias (multipath) and spatial decorrelation of the atmosphere.

Phase-Dominated DGNSS (RTK or PPP-RTK) Positioning

For the carrier-phase observations, we can apply a similar technique as for code. The expectation of the (nonlinear) phase-observation equation reads (21.2)

$$\begin{aligned} E(\varphi_{r,j}^s(t)) &= \rho_r^s(t, t - \tau_r^s) + T_r^s(t) - \mu_j I_r^s(t) \\ &\quad + c[dt_r(t) + \delta_{r,j}] + \lambda_j N_{r,j}^s \\ &\quad - c[dt^s(t - \tau_r^s) - \delta_j^s]. \end{aligned} \quad (21.81)$$

For a reference receiver 1, we can subtract the observed carrier-phase from the computed range, which yields the *phase-range corrections* (PRC) for satellite s

$$\text{PRC}_{\varphi,j}^s(t) = \rho_1^s(t, t - \tau_1^s) - \varphi_{1,j}^s(t). \quad (21.82)$$

These phase-range corrections are in a next step applied to correct the carrier-phases of the user r , corresponding to the same satellite s

$$\begin{aligned} E(\tilde{\varphi}_{r,j}^s(t)) &= \rho_r^s(t, t - \tau_r^s) + T_{1r}^s(t) - \mu_j I_{1r}^s(t) \\ &\quad + c[dt_{1r}(t) + \delta_{1r,j}] + \lambda_j N_{1r,j}^s. \end{aligned} \quad (21.83)$$

Here the corrected carrier-phase reads $\tilde{\varphi}_{r,j}^s(t) = \varphi_{r,j}^s(t) + \text{PRC}_{\varphi,j}^s(t)$. If the latency of the corrections equals zero, DGNSS positioning with phase (and code) is better known as the *real-time kinematic* (RTK) positioning technique. The accuracy of RTK positioning is at centimeter level (or better), *provided* that the carrier-phase ambiguities can be resolved to their integer values. We remark that the corrected phase-observation equations cannot be directly used for positioning as the system is (as with PPP) rank deficient. How to deal with this is discussed in Sect. 21.4.4.

In practice, under average ionospheric conditions, the differential atmospheric errors can be neglected for distances between the receivers up to about 10 km, and the positioning method is referred to as *short-baseline RTK*. Satellite orbit errors can also be ignored for these short distances (Sect. 21.4.2). RTK based on GPS is a proven positioning concept and when GPS is combined with other constellations even more promising results are obtained (e.g., [21.53] for RTK based on GPS with Galileo and [21.54] for GPS combined

with BeiDou). Extending the RTK technique to longer distances is possible, but then the differential atmospheric errors and orbit errors need to be taken into account (long-baseline RTK) [21.55, 56]. In case of *network RTK* corrections for the atmospheric delays are estimated from a network of surrounding reference stations and transmitted to users. More details about (network) RTK positioning can be found in Chap. 26. The technique of *PPP-RTK* [21.57] (or PPP-AR; AR = Ambiguity Resolution) also relies on correction information provided by a reference network, but the main difference with network RTK is that with PPP-RTK the correction are provided in the *parameter* space, while in case of network RTK the corrections are in the *observation* space. PPP-RTK is discussed in further detail in Sect. 21.4.5.

21.4.2 Impact of Orbit Errors

When linearizing the GNSS observation equations, the satellite positions are held fixed to their values as computed (per receiver) in the SPP processing. The accuracy of satellite positions computed from the broadcast ephemeris is at the level of a few meters, while based on precise ephemeris this is (for new constellations: expected) at the level of 5–10 cm [21.58]. Errors in these fixed satellite positions may negatively impact the estimated receiver position. If we denote the errors in the satellite position as vector $\Delta \mathbf{r}^s$, then its effect on the relative baseline is upper bounded according to the following rule-of-thumb [21.59]

$$\left| [\mathbf{e}_{1r}^s(t)]^\top \Delta \mathbf{r}^s(t) \right| \leq \frac{\|\mathbf{r}_{1r}(t)\|}{\|\mathbf{r}_r^s(t)\|} \|\Delta \mathbf{r}^s(t)\|, \quad (21.84)$$

with $\mathbf{e}_{1r}^s(t) = \mathbf{e}_r^s(t) - \mathbf{e}_1^s(t)$ the between-receiver differenced LOS vector, $\mathbf{r}_{1r}(t) = \mathbf{r}_r(t) - \mathbf{r}_1(t)$ the relative receiver position (baseline) vector, and $\mathbf{r}_r^s(t) = \mathbf{r}^s(t) - \mathbf{r}_r(t)$ the receiver–satellite position vector. For example, the impact of an orbit error of 2 m on a baseline of 100 km is at most only 1 cm (assuming a 20 000 km receiver–satellite range). The above upper bound can also be used to assess the effect of differences in reference frames of the satellite positions in a multiconstellation case.

For longer baselines (and network-RTK or PPP-RTK) precise ephemeris should be used, restricting the impact of orbit errors on the receiver position.

21.4.3 Ionosphere-Fixed/Weighted/Float Models

Differential ionospheric delays generally become larger for increasing baseline lengths. To flexibly apply the

models to a whole range of baseline lengths, the relative models presented in this section are presented for three different versions regarding the presence of the differential ionospheric delays.

For sufficiently short baselines the *ionosphere-fixed* model is presented, in which the differential ionospheric delays are absent. This corresponds to the assumption that the absolute ionospheric delays for a certain satellite are equal for all receivers, as they intersect the ionosphere in the same part, that is, $I_1^s(t) = I_r^s(t) \doteq I^s(t)$ (Fig. 21.7). For longer baselines the size of the differential ionospheric delays may be within certain bounds such that knowledge can be incorporated in the form of (soft) constraints. The resulting model is referred to as the *ionosphere-weighted* model. For even longer baselines, when we do not have any a-priori knowledge on the differential ionospheric delays, the ionospheric delays are assumed as completely unknown parameters. This is the *ionosphere-float* model. This terminology concerning the tuning of the ionospheric delays is adopted from [21.60].

21.4.4 Undifferenced Relative Positioning Models

In this subsection, the relative positioning model is presented based on the original *undifferenced, uncombined* observation equations for code and phase. The link with the models based on differenced observations will be

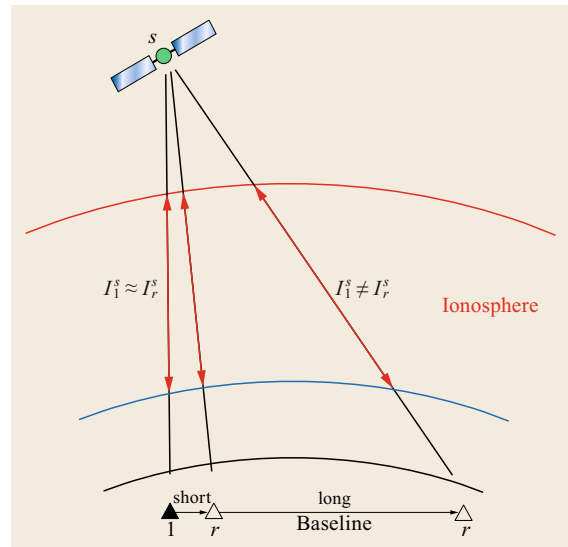


Fig. 21.7 Visualization of propagation of GNSS signals through the ionosphere. Signals of receivers that are relatively close to each other travel through similar parts of the ionosphere, while more remote receivers experience a more different ionospheric delay

made later on in this chapter. An important advantage of the undifferenced model is that it is more flexible than differenced; this flexibility has been recognized already for a long time [21.18, 61–64]. For example, satellites that are only visible by some of the receivers in the network can still be used, while differencing algorithms can only process those satellites that are in view by *all* receivers. Another important advantage of the undifferenced model is that temporal constraints can be incorporated to parameters that would have been eliminated when differencing, for example, clocks or hardware delays. The undifferenced models that are presented in this section are assumed to be valid for a *network* of n GNSS receivers. By simply setting $n = 2$ the results for a *single-baseline* model are obtained.

Rank-Deficient Undifferenced Model

As with the undifferenced models for point positioning (Sect. 21.3), also for the relative positioning model based on undifferenced observations it is not possible to estimate all parameters uniquely because the system of observation equations is rank deficient. To overcome this rank deficiency, as in the case of the point positioning models, we apply the theory of S-systems (Chap. 22), resulting in linear combinations of parameters that are estimable. There is, however, not a unique way to choose the estimable linear combinations; in theory there are infinite possibilities. Different choices lead to different *interpretations* of the estimable parameters.

One choice leads to the so-called *distinct clocks* model, for which a receiver clock as well as a satellite clock parameter becomes estimable for each frequency for all code and phase data [21.18, 65]. Another choice results in a full-rank undifferenced model in which common receiver clock and satellite clock parameters become estimable. This is the so-called *common clocks* model [21.66, 67]. This model is reviewed here, as its estimable parameters have a clear link with those presented in the previous sections for SPP and PPP. The common clocks model as presented here applies to a general case of $f \geq 2$ frequencies.

Regional Networks

For distances between the receivers that are smaller than about 500 km usually not the position and ZTD of all receivers are estimated in absolute sense by means of relative models, as the LOS vectors of the different receivers with respect to the same satellite become close to parallel, resulting in poorly estimable absolute positions and ZTDs [21.68]. In the extreme case, if the LOS vectors are assumed to be equivalent for the different receivers, that is, $\mathbf{g}_1^s(t) = \dots = \mathbf{g}_n^s(t) \doteq \mathbf{g}^s(t)$ (21.72), this causes an additional rank deficiency in the network model. A common procedure to overcome this addi-

tional rank deficiency is to estimate the position and ZTD of all receivers relative to that of one of receivers (i. e., the so-called *pivot receiver*).

Common Clocks Undifferenced Model

The estimable parameter functions corresponding to the common clocks model have been derived in [21.69] and are presented in Table 21.4. To differentiate the estimable parameters from their original counterparts, the estimable ones are denoted using a *tilde*. Concerning the interpretation, it can first of all be seen that all estimable parameters are a function of their original parameter, but biased by one or more other parameters. These other parameters are frequently the parameters corresponding to the pivot receiver and/or pivot satellite (for both we selected the first receiver and first satellite, denoted using a 1 subscript or superscript, but this could be any other receiver and satellite in the network).

The estimable receiver/ZTD parameters ($\tilde{\mathbf{x}}_r(t)$) are absolutely estimable for global networks and relatively estimable in case of a regional network. Absolutely is put between quotes here, as the receiver positions are still relative with respect to the satellite positions that are held fixed. The presence of the pivot receiver's position+ZTD is however *compensated* by their presence in the estimable satellite clock ($\tilde{d}^s_r(t)$) from a regional network. This compensation holds for all biases present in a certain estimable parameter. Note that the estimable satellite clock can be written as a combination of ionosphere-free satellite clock, minus the ionosphere-free clock of the pivot receiver, plus (in case of a regional network) the pivot receiver's position and/or ZTD.

If the ionosphere is assumed to be *float* instead of fixed/weighted, this causes an additional rank deficiency, which leads to differences in interpretation of the estimable receiver clock ($\tilde{d}^s_r(t)$), receiver phase and code delay ($\tilde{\delta}_{r,j}$ and $\tilde{a}_{r,j}$, respectively), as well as the estimable ionospheric delay ($\tilde{I}^s_r(t)$). Note the subtle difference in the interpretation of the estimable ionospheric delay between the ionosphere-weighted and ionosphere-float models: in the first case the DCB of the pivot receiver appears (i. e., $\text{DCB}_{1,12}$), while in the second case it is the DCB corresponding to the receiver for which the ionospheric delay parameter is considered (i. e., $\text{DCB}_{r,12}$). Furthermore, note that the interpretation of the ionospheric delay in case the ionosphere is float is exactly the same as the estimable ionospheric delay parameter of the dual-frequency PPP model in Table 21.3, if no satellite DCB corrections are applied.

Special attention needs the estimable satellite code delay parameter (\tilde{d}^s_j). From Table 21.4 it follows that it is estimable as a modernized DCB (i. e., frequency j relative to frequency 1), however, with respect to the (traditional) DCB between the first two frequencies.

Table 21.4 Estimable undifferenced parameters for the common clocks relative positioning model. Note: the (modernized) receiver DCB is defined as $\text{DCB}_{r,1j}(t) = d_{r,1}(t) - d_{r,j}(t)$, while the (modernized) satellite DCB is defined as $\text{DCB}_{1j}^s(t) = d_1^s(t) - d_j^s(t)$

Estimable parameter	Notation and interpretation	Conditions
Receiver position/ZTD	$\tilde{\mathbf{x}}_r(t) = \mathbf{x}_r(t) - \underbrace{\mathbf{x}_1(t)}_{\text{if regional network}}$	$r \geq 1$ ($r \geq 2$ reg. net)
Receiver clock	$d\tilde{t}_r(t) = [d_{r,1}(t) + d_{r,1}] - [d_{1,1}(t) + d_{1,1}] + \underbrace{\frac{\mu_1}{\mu_2 - \mu_1} [\text{DCB}_{r,12} - \text{DCB}_{1,12}]}_{\text{if ionosphere float}}$	$r \geq 2$
Receiver-phase delay	$\tilde{\delta}_{r,j} = \left[\delta_{r,j} - d_{r,1} + \frac{\lambda_j}{c} N_{r,j}^1 \right] - \left[\delta_{1,j} - d_{1,1} + \frac{\lambda_j}{c} N_{1,j}^1 \right] - \underbrace{\frac{\mu_j + \mu_1}{\mu_2 - \mu_1} [\text{DCB}_{r,12} - \text{DCB}_{1,12}]}_{\text{if ionosphere float}}$	$r \geq 2, j \geq 1$
Receiver-code delay	$\tilde{d}_{r,j} = -[\text{DCB}_{r,1j} - \text{DCB}_{1,1j}] + \underbrace{\frac{\mu_j - \mu_1}{\mu_2 - \mu_1} [\text{DCB}_{r,12} - \text{DCB}_{1,12}]}_{\text{if ionosphere float}}$	$r \geq 2, j \geq 2$ ($j \geq 3$ iono float)
Satellite clock	$d\tilde{t}^s(t) = [d_1^s(t) - d_{1F}^s] - [d_{1,1F}(t) + d_{1,1F}] - \underbrace{[\mathbf{g}^s(t)]^\top \mathbf{x}_1(t)}_{\text{if regional network}}$	$s \geq 1$
Satellite-phase delay	$\tilde{\delta}_j^s = \left[\delta_j^s - d_{1F}^s - \frac{\mu_j}{\mu_2 - \mu_1} \text{DCB}_{12}^s + \frac{\lambda_j}{c} N_{1,j}^s \right] + \left[\delta_{1,j} - d_{1,1F} - \frac{\mu_j}{\mu_2 - \mu_1} \text{DCB}_{1,12} \right]$	$s \geq 1, j \geq 1$
Satellite-code delay	$\tilde{d}_j^s = -[\text{DCB}_{1j}^s - \frac{\mu_j - \mu_1}{\mu_2 - \mu_1} \text{DCB}_{12}^s] - [\text{DCB}_{1,1j} - \frac{\mu_j - \mu_1}{\mu_2 - \mu_1} \text{DCB}_{1,12}]$	$s \geq 1, j \geq 3$
Ionospheric delay	$\tilde{I}_r^s(t) = \begin{cases} I_r^s(t) - \frac{1}{\mu_2 - \mu_1} c [\text{DCB}_{12}^s + \text{DCB}_{1,12}] & \text{if ionosphere-fixed} \\ I_r^s(t) - \frac{1}{\mu_2 - \mu_1} c [\text{DCB}_{12}^s + \text{DCB}_{1,12}] & \text{if ionosphere-weighted } (r \geq 1) \\ I_r^s(t) - \frac{1}{\mu_2 - \mu_1} c [\text{DCB}_{12}^s + \text{DCB}_{r,12}] & \text{if ionosphere-float } (r \geq 1) \end{cases}$	$s \geq 1$
Phase ambiguities	$\tilde{N}_{r,j}^s = [N_{r,j}^s - N_{1,j}^s] - [N_{r,j}^1 - N_{1,j}^1]$	$r \geq 2, s \geq 2, j \geq 1$

This whole satellite-dependent DCB term is also relative to a similar DCB term, but then for the pivot receiver. As a consequence, satellite code parameters are only estimable when *at least three frequencies* are used (i.e., if $j \geq 3$, thus for example not in the legacy dual-frequency GPS case).

Concerning the phase ambiguities ($\tilde{N}_{r,j}^s$), they are estimable as *double-differenced* (DD) parameters, and thus integer valued, with respect to the network's pivot receiver and pivot satellite. By means of integer ambiguity resolution (Chap. 23), the network parameters can be estimated with the highest possible precision.

Based on the estimable undifferenced parameter functions, the system of *full-rank* GNSS network observation equations can be given as follows

$$\begin{aligned}
 E(\Delta p_{r,j}^s(t)) &= \mathbf{g}_{[r]}^s(t)^\top \tilde{\mathbf{x}}_r(t) + c[d\tilde{t}_r(t) + \tilde{d}_{r,j}] \\
 &\quad - c[d\tilde{t}^s(t) - \tilde{d}_j^s] + \mu_j \tilde{I}_r^s(t), \quad r \geq 1, \\
 E(\Delta \varphi_{r,j}^s(t)) &= \mathbf{g}_{[r]}^s(t)^\top \tilde{\mathbf{x}}_r(t) + c[d\tilde{t}_r(t) + \tilde{\delta}_{r,j}] \\
 &\quad - c[d\tilde{t}^s(t) - \tilde{\delta}_j^s] - \mu_j \tilde{I}_r^s(t) + \lambda_j \tilde{N}_{r,j}^s, \\
 r \geq 1, \quad [E(I_r^s(t) - I_1^s(t)) &= \tilde{I}_r^s(t) - \tilde{I}_1^s(t), r \geq 2].
 \end{aligned}
 \tag{21.85}$$

Note that the between-receiver ionospheric constraints, which are included in the form of pseudo observation equations, only appear in case the ionosphere is weighted (that is why we denote them using square brackets).

The *redundancy* of the undifferenced network model reads in the ionosphere-fixed/weighted cases $(n-1)f(m-1) - 4n + (k-1)[n(2fm-1) - (2m-1) - 4n]$, where the 4 reflects the estimation of both positions and ZTD parameters in the network. In case of a regional network, the $4n$ is to be replaced by $4(n-1)$. In the ionosphere-float case it reads $(n-1)(f-1)(m-1) - 4n + (k-1)[n(\{2f-1\}m-1) - (m-1) - 4n]$, requiring one more frequency.

21.4.5 PPP-RTK Models

As discussed in Sect. 21.3.7, (standard) PPP is possible by applying satellite positions and clocks that are determined by a (global) reference network (and satellite DCBs in case of single-frequency PPP). This information is not sufficient for the single-receiver GNSS user to resolve the ambiguities in his carrier-phase observations to integer values, which is needed for high-precision PPP based on short convergence times.

The information that is lacking for integer ambiguity resolution enabled PPP (PPP-RTK or PPP-AR) are corrections for satellite phase and code biases, since these parameters hamper the estimable PPP ambiguities from being integer (Table 21.3).

If the reference network adopts an undifferenced model formulation, the crucial satellite phase and code bias information is among the estimable parameters. We emphasize that the choice of a common clocks network model (as done in the previous subsection) is not a prerequisite for PPP-RTK; the network may also adopt another S-system. In fact, the user does not even need to know the S-system of the network as he can equally apply corrections determined by networks that are based on different S-systems [21.69]. However, if the reference network adopts the common clocks model, the estimable satellite clocks are ionosphere-free parameters and this allows a direct comparison with the clocks of, for example, the IGS.

In the literature, other network models can be found that serve as basis for the generation of PPP-RTK corrections [21.35, 70–72]. Usually these network models are not based on strictly undifferenced observables, but on linear combinations between observables, of which the *ionosphere-free* and *Melbourne–Wübbena* combinations (Chap. 20) are frequently used. Consequently the hardware delay and ambiguity parameters estimated using these linear combinations of observables are in the form of *wide-lane* and *narrow-lane* combinations. One-to-one transformation formulas between the corrections of various PPP-RTK approaches are presented in [21.73].

As with the discussion of the models for PPP, in the following we make a distinction between PPP-RTK models without ionospheric corrections (ionosphere-float PPP-RTK model) and those including ionospheric corrections (ionosphere-corrected PPP-RTK model).

Ionosphere-Float PPP-RTK Model

In the absence of ionospheric corrections, the PPP-RTK corrections for code and phase are the satellite clock as well as satellite code and phase-delay parameters from the common clocks network (and tropospheric and other corrections)

$$\begin{aligned} o_{p_{r,j}}^s(t) &= c \left[d\tilde{r}^s(t) - \tilde{d}_j^s \right] - T_{r,0}^s(t), \\ o_{\varphi_{r,j}}^s(t) &= c \left[d\tilde{r}^s(t) - \tilde{\delta}_j^s \right] - T_{r,0}^s(t). \end{aligned} \quad (21.86)$$

Here \tilde{d}_j^s is only applied in case $j \geq 3$, that is, for code observations at a third or higher frequency. Thus, in the legacy dual-frequency GPS case it is not applied, but it is for the third frequency observations transmitted by the Block IIF GPS satellites.

If the PPP-RTK user adopts – like the network – a common clocks model, the estimable user parameters plus their interpretation automatically follow from Table 21.4 by regarding receiver r to be the user receiver, instead of a network receiver. The difference is that there are no satellite clock as well as satellite phase and code bias parameters for the user, as these are corrected for. However, the interpretation of the estimable user's position+ZTD, receiver clock, phase delay, code delay, ionospheric delay, and ambiguity parameters is as given in the table, for the case the ionosphere is float. The only difference in the interpretation of the user parameters may be the pivot satellite, showing up in the estimable phase delays and ambiguities. It is namely not needed that the user should adopt the *same* pivot satellite as the network; this can be any of the satellites he has in view.

For the PPP-RTK user's position and ZTD it means that in case of a regional network they are – like for the network receivers – estimable relative to those of the pivot receiver in the network. In case the network pivot receiver's position is held fixed in the processing (a usual assumption in case of CORS networks, i.e., $\Delta r_1(t) = 0$), the estimable user's position is not relative to the network's pivot receiver; however the user's estimable ZTD parameter still is, that is,

$$\tilde{T}_r^z(t) = T_r^z(t) - T_1^z(t).$$

The estimable PPP-RTK user's ambiguity is also relative, with respect to the network pivot receiver's ambiguity and with respect to an arbitrarily chosen pivot satellite p

$$\tilde{N}_{r,j}^s = [N_{r,j}^s - N_{1,j}^s] - [N_{r,j}^p - N_{1,j}^p], \quad (21.87)$$

for $s \neq p$ and $j \geq 1$. It is estimable as a double-differenced ambiguity and thus standard ambiguity resolution (LAMBDA) is applied to estimate the integer PPP-RTK ambiguities. Usually a three-step-procedure is followed to solve the position based on the integer ambiguities (Chap. 23).

If the (corrected) pseudorange and carrier-phase observables of, in general, f frequencies are denoted as vectors $\Delta \tilde{\mathbf{p}}_r(t)$ and $\Delta \tilde{\boldsymbol{\varphi}}_r(t)$, respectively, the full-rank undifferenced, multifrequency PPP-RTK model is given in Table 21.5 (*second row*). It is emphasized that the receiver code bias parameters ($\tilde{d}_{r,j}$) only show up in case of triple- or higher frequency (i.e., $j \geq 3$) and are absent in a dual-frequency case.

Ionosphere-Corrected PPP-RTK Model

The ionosphere-float PPP-RTK approach of which the model was presented in the previous section requires

Table 21.5 Full-rank undifferenced multifrequency (MF) PPP-RTK modelsMF PPP-RTK (ionosphere-corrected; $f \geq 1$):

$$E \begin{pmatrix} \Delta \tilde{\mathbf{p}}_{r,1}(t) \\ \Delta \tilde{\mathbf{p}}_{r,2}(t) \\ \vdots \\ \Delta \tilde{\mathbf{p}}_{r,f}(t) \\ \hline \Delta \tilde{\boldsymbol{\varphi}}_{r,1}(t) \\ \vdots \\ \Delta \tilde{\boldsymbol{\varphi}}_{r,f}(t) \end{pmatrix} = \begin{bmatrix} \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{u}_m & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \dots & \mathbf{u}_m & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \hline \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \dots & \mathbf{0} & \mathbf{u}_m & \dots & \mathbf{0} & \lambda_1 \mathbf{C}_m & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{u}_m & \mathbf{0} & \dots & \lambda_f \mathbf{C}_m \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_r(t) \\ cd\tilde{t}_r(t) \\ \hline \tilde{cd}_{r,2} \\ \vdots \\ \tilde{cd}_{r,f} \\ \hline \tilde{cd}_{r,1} \\ \vdots \\ \tilde{cd}_{r,f} \\ \hline \tilde{N}_{r,1} \\ \vdots \\ \tilde{N}_{r,f} \end{bmatrix}$$

MF PPP-RTK (ionosphere-float; $f \geq 2$):

$$E \begin{pmatrix} \Delta \tilde{\mathbf{p}}_{r,1}(t) \\ \Delta \tilde{\mathbf{p}}_{r,2}(t) \\ \Delta \tilde{\mathbf{p}}_{r,3}(t) \\ \vdots \\ \Delta \tilde{\mathbf{p}}_{r,f}(t) \\ \hline \Delta \boldsymbol{\varphi}_{r,1}(t) \\ \vdots \\ \Delta \boldsymbol{\varphi}_{r,f}(t) \end{pmatrix} = \begin{bmatrix} \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mu_1 \mathbf{I}_m & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mu_2 \mathbf{I}_m & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{u}_m & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mu_3 \mathbf{I}_m & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \dots & \mathbf{u}_m & \mathbf{0} & \dots & \mathbf{0} & \mu_f \mathbf{I}_m & \mathbf{0} & \dots & \mathbf{0} \\ \hline \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \dots & \mathbf{0} & \mathbf{u}_m & \dots & \mathbf{0} & -\mu_1 \mathbf{I}_m & \lambda_1 \mathbf{C}_m & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_r(t) & \mathbf{u}_m & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{u}_m & -\mu_f \mathbf{I}_m & \mathbf{0} & \dots & \lambda_f \mathbf{C}_m \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_r(t) \\ cd\tilde{t}_r(t) \\ \hline \tilde{cd}_{r,3} \\ \vdots \\ \tilde{cd}_{r,f} \\ \hline \tilde{cd}_{r,1} \\ \vdots \\ \tilde{cd}_{r,f} \\ \hline \tilde{\mathbf{I}}_r(t) \\ \tilde{N}_{r,1} \\ \vdots \\ \tilde{N}_{r,f} \end{bmatrix}$$

a relatively long time before the ambiguities have converged and the integers can be resolved. To speed this up, it is essential to incorporate ionospheric corrections. Global reference networks (such as the IGS network) provide ionospheric corrections in the form of GIMs. Although these global ionospheric corrections may serve standard PPP, for PPP-RTK, aiming at centimeter level precision, they are not precise enough. More precise ionospheric corrections can be generated by a *regional* reference network that better captures the spatial variation of the ionospheric delays than a global network.

A way to generate these regional ionospheric corrections is by means of *Kriging interpolation* [21.74] of the estimated ionospheric delays at the network receivers to the approximate location of the user [21.67]. The interpolation, carried out on a satellite-by-satellite

basis, can be given as

$$\mathbf{I}_r^s(t) = \mathbf{h}_r^\top [\mathbf{I}_1^s(t), \dots, \mathbf{I}_n^s(t)]^\top. \quad (21.88)$$

Here \mathbf{h}_r denotes the n -vector performing the interpolation over the n network receivers and $\mathbf{I}_r^s(t)$ denotes the interpolated ionospheric delay at the (approximate) location of the user. The entries of the interpolation vector depend on the assumed spatial coherence of the ionosphere, as well as on the distances of the PPP-RTK user with respect to the network receivers. A property of the Kriging interpolation vector is that its entries add up to 1 (i.e., $\mathbf{h}_r^\top \mathbf{u}_n = 1$). At first sight there seems to be a problem to perform the interpolation, as it is based on the original, unbiased ionospheric delays, while the ionospheric parameters that are estimable are biased by other parameters. Fortunately this is not a prob-

lem, and the interpolation should simply be based on the estimable biased ionospheric parameters. If the network model is in the ionosphere-float common clocks S-system, the interpolation of the estimated network ionospheric corrections can then be written as

$$\tilde{I}_r^s(t) = I_r^s(t) - \frac{1}{\mu_2 - \mu_1} c [\text{DCB}_{12}^s + \text{DCB}_{\bar{r},12}] . \quad (21.89)$$

Thus, the interpolated ionospheric correction the PPP-RTK user should apply can be interpreted as the interpolated ionospheric delay itself, plus the satellite DCB (the interpolation does not affect this since it is the same for all receivers), minus the interpolated receiver DCB, that is,

$$\text{DCB}_{\bar{r},12} = \mathbf{h}_{\bar{r}}^T [\text{DCB}_{1,12}, \dots, \text{DCB}_{n,12}]^T .$$

The presence of the satellite DCB as a bias of the interpolated ionospheric corrections means that the user does not have to explicitly correct for it, as is the case when, for example, GIM-based ionospheric corrections are used.

The PPP-RTK corrections for code and phase can now be given as

$$\begin{aligned} o_{r,j}^s(t) &= c [d\tilde{r}^s(t) - \tilde{d}_j^s] - T_{r,0}^s(t) - \mu_j \tilde{I}_r^s(t) , \\ o_{\varphi_{r,j}}^s(t) &= c [d\tilde{r}^s(t) - \tilde{\delta}_j^s] - T_{r,0}^s(t) + \mu_j \tilde{I}_r^s(t) . \end{aligned} \quad (21.90)$$

It is assumed that the expectation of the user's predicted ionospheric delay corresponds to its true ionospheric delay, that is,

$$E(I_r^s(t)) = E(I_r^s(t)) .$$

Like with the ionosphere-float PPP-RTK model, if in the ionosphere-corrected case the user adopts a common clocks model as well, his estimable position/ZTD, receiver clock, receiver hardware bias, and ambiguity parameters follow from Table 21.4 by regarding receiver r to be the user receiver. Important to emphasize is that the *ionosphere-float* DCB parameters within the estimable receiver clock as well as receiver phase/code

delay apply in this case as well, with the difference that $\text{DCB}_{r,12}$ should be replaced by its network-interpolated counterpart $\text{DCB}_{\bar{r},12}$. Important consequence is that the estimable receiver code DCB is already estimable for two frequencies (i. e., $j \geq 2$) instead of three.

The PPP-RTK model for the ionosphere-corrected case is given in Table 21.5 (*first row*). Note that as the ionospheric delays are corrected for there are no ionospheric parameters. Furthermore, the receiver code delays are estimable from the second frequency onward, in contrast to the ionosphere-float PPP-RTK model, where they are estimable only from the third frequency and higher. In the single-frequency ($f = 1$) case, the model reduces to the PPP model given in Table 21.3 (SF-PPP ionospheric-correction).

21.4.6 Link Between PPP-RTK and PPP

If the full-rank PPP-RTK design matrix in Table 21.5 is considered for two frequencies (i. e., $f = 2$), such that the part for the receiver code biases disappears, it is exactly *identical* to the full-rank design matrix corresponding to dual-frequency standard PPP (Table 21.3) (*DF PPP iono-float*). Although the interpretation of the parameters differs between PPP and PPP-RTK, the solution of PPP and PPP-RTK with the ambiguities treated as float are identical. Hence, standard PPP can be considered as a *special case* of PPP-RTK. In relation to this, it follows from the interpretations that the satellite phase biases, that is, $\tilde{\delta}_j^s$, exactly correct for the bias that is inside the PPP ambiguities in order to make them become PPP-RTK ambiguities and thus integer. Thus, in general

$$[\tilde{N}_{r,j}^s]_{\text{PPP}} - \frac{c}{\lambda_j} [\tilde{\delta}_j^s - \tilde{\delta}_j^p] = [\tilde{N}_{r,j}^s]_{\text{PPP-RTK}} . \quad (21.91)$$

Here $[\tilde{N}_{r,j}^s]_{\text{PPP}}$ denotes the estimable PPP ambiguity of which its interpretation is given in Table 21.3 and $[\tilde{N}_{r,j}^s]_{\text{PPP-RTK}}$ denotes the estimable integer PPP-RTK ambiguity as in (21.87). Thus, when the estimable between-satellite differenced satellite phase bias is subtracted from the estimable PPP ambiguities, the estimable integer PPP-RTK ambiguities are obtained.

21.5 Differenced Positioning Models

Differencing techniques are traditionally applied in GNSS processing to reduce the amount of unknowns and observations. However, they may also result in a loss of information, for example, in the multiepoch case incorporating temporal constraints on parameters which would otherwise be eliminated by means of differencing. This section briefly presents the differenced versions of the single-constellation positioning models presented earlier in this chapter. Although not discussed in detail, we mention that differencing causes (mathematical) correlation between the differenced observations and this should be appropriately taken into account through the variance–covariance matrix of the observations.

21.5.1 Single Differencing

Differencing the observations with respect to a chosen pivot satellite removes the receiver-dependent parameters from the models (Fig. 21.8a). This section presents the between-satellite differenced versions of the SPP, PPP, and PPP-RTK models.

Between-Satellite Differenced SPP Model

Taking the differences of the code observables between satellite s and (pivot) satellite p , the single-differenced, single-frequency SPP model (21.50) becomes

$$E(\mathbf{D}_m^\top \Delta \tilde{\mathbf{p}}_{r,j}(t)) = [\mathbf{D}_m^\top \mathbf{G}_r(t)] \Delta \mathbf{r}_r(t). \quad (21.92)$$

Here the $(m-1) \times m$ (transposed) *differencing matrix* is defined as

$$\mathbf{D}_m^\top = \begin{bmatrix} \mathbf{I}_{p-1} & -\mathbf{u}_{p-1} & \mathbf{0} \\ \mathbf{0} & -\mathbf{u}_{m-p} & \mathbf{I}_{m-p} \end{bmatrix}. \quad (21.93)$$

A property of this differencing matrix is that $\mathbf{D}_m^\top \mathbf{u}_m = \mathbf{0}$. In the *dual-frequency* case in which atmospheric delays are estimated, see (21.53) for the undifferenced dual-frequency SPP model, its between-satellite differenced counterpart becomes

$$\begin{aligned} E \left(\begin{bmatrix} \mathbf{D}_m^\top \Delta \tilde{\mathbf{p}}_{r,1}(t) \\ \mathbf{D}_m^\top \Delta \tilde{\mathbf{p}}_{r,2}(t) \end{bmatrix} \right) \\ = \begin{bmatrix} \mathbf{D}_m^\top \mathbf{G}_r(t) & \mu_1 \mathbf{I}_{m-1} \\ \mathbf{D}_m^\top \mathbf{G}_r(t) & \mu_2 \mathbf{I}_{m-1} \end{bmatrix} \begin{bmatrix} \Delta \mathbf{r}_r(t) \\ \mathbf{D}_m^\top \tilde{\mathbf{I}}_r(t) \end{bmatrix}. \end{aligned} \quad (21.94)$$

The between-satellite differenced ionospheric delay, which is denoted as vector $\mathbf{D}_m^\top \tilde{\mathbf{I}}_r(t)$, is free of the receiver DCB term, which appears in the undifferenced case (21.54).

Between-Satellite Differenced Multiconstellation SPP Model

In the presence of observations of two constellations, the SPP model in its undifferenced form was given in (21.59), assuming one frequency per constellation. In this case, the between-satellite differencing can be carried out in different ways. A first way is to choose a pivot satellite for each constellation and difference the observations corresponding to its own constellation-specific pivot satellite. A second way is to difference the observations of both constellations to one *common* pivot satellite.

In the first case, using a pivot satellite for each constellation, the between-satellite differenced model reads simply

$$E \left(\begin{bmatrix} \mathbf{D}_{m_A}^\top \Delta \tilde{\mathbf{p}}_{r,j}^A(t) \\ \mathbf{D}_{m_B}^\top \Delta \tilde{\mathbf{p}}_{r,j}^B(t) \end{bmatrix} \right) = \begin{bmatrix} \mathbf{D}_{m_A}^\top \mathbf{G}_r^A(t) \\ \mathbf{D}_{m_B}^\top \mathbf{G}_r^B(t) \end{bmatrix} \Delta \mathbf{r}_r(t). \quad (21.95)$$

Here m_A and m_B denote the number of satellites for constellation A and B, respectively. Compared to its undifferenced counterpart in (21.59), the receiver clock common for both constellations, as well as the ISB parameter for the observations of constellation B, have been eliminated. In the second case, where the observations of both are differenced with respect to the pivot satellite selected from constellation A, the between-satellite differenced model reads

$$\begin{aligned} E \left(\mathbf{D}_{m_A+m_B}^\top \begin{bmatrix} \Delta \tilde{\mathbf{p}}_{r,j}^A(t) \\ \Delta \tilde{\mathbf{p}}_{r,j}^B(t) \end{bmatrix} \right) \\ = \left[\mathbf{D}_{m_A+m_B}^\top \begin{pmatrix} \mathbf{G}_r^A(t) \\ \mathbf{G}_r^B(t) \end{pmatrix} \right] \begin{pmatrix} \mathbf{0} \\ \mathbf{u}_{m_B} \end{pmatrix} \begin{bmatrix} \Delta \mathbf{r}_r(t) \\ c \text{ISB}_{r,j}^{\text{AB}} \end{bmatrix}. \end{aligned} \quad (21.96)$$

Here $\mathbf{D}_{m_A+m_B}^\top$ denotes the $(m_A+m_B-1) \times (m_A+m_B)$ difference matrix. Due to the differencing between constellations the ISB parameter is not eliminated. Both models (21.95) and (21.96) are however equivalent in terms of redundancy and positioning solution. Although model (21.96) has one parameter more than model (21.95), it has also one more observation. However, the situation changes if the ISB can be assumed known. In that case, the observations of constellation B can be corrected for it such that model (21.96) reduces to

$$\begin{aligned} E \left(\mathbf{D}_{m_A+m_B}^\top \begin{bmatrix} \Delta \tilde{\mathbf{p}}_{r,j}^A(t) \\ \Delta \tilde{\mathbf{p}}_{r,j}^B(t)' \end{bmatrix} \right) \\ = \begin{bmatrix} \mathbf{D}_{m_A+m_B}^\top \begin{pmatrix} \mathbf{G}_r^A(t) \\ \mathbf{G}_r^B(t) \end{pmatrix} \end{bmatrix} \Delta \mathbf{r}_r(t). \end{aligned} \quad (21.97)$$

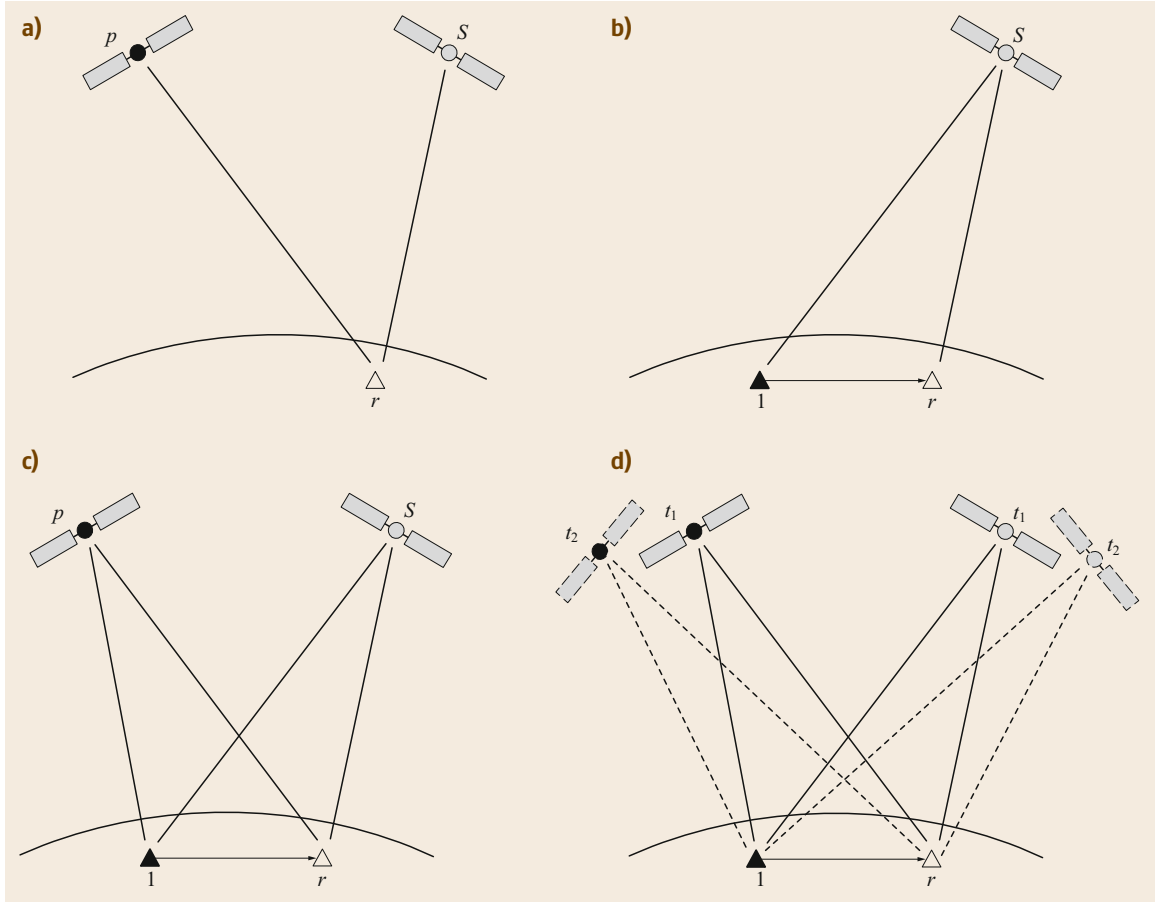


Fig. 21.8a–d Various differencing strategies: **(a)** between-satellite single differencing; **(b)** between-receiver single differencing; **(c)** double differencing; **(d)** triple differencing

As a consequence, the above ISB-corrected model can be considered as a *single-constellation model*, similar to (21.92), but now with $m_A + m_B$ satellites.

Between-Satellite Differenced PPP(–RTK) Model

In case of single-constellation, single-frequency (ionosphere-corrected) PPP(–RTK), between-satellite differencing results in the following observation (Table 21.3)

$$\begin{aligned} E \left(\begin{bmatrix} \mathbf{D}_m^\top \Delta \tilde{\mathbf{p}}_{r,j}(t) \\ \mathbf{D}_m^\top \Delta \tilde{\boldsymbol{\phi}}_{r,j}(t) \end{bmatrix} \right) \\ = \begin{bmatrix} \mathbf{D}_m^\top \mathbf{G}_r(t) & \mathbf{0} \\ \mathbf{D}_m^\top \mathbf{G}_r(t) & \lambda_j \mathbf{I}_{m-1} \end{bmatrix} \begin{bmatrix} \mathbf{x}_r(t) \\ \tilde{\mathbf{N}}_{r,j} \end{bmatrix}. \end{aligned} \quad (21.98)$$

Here use is made of the property that $\mathbf{D}_m^\top \mathbf{C}_m = \mathbf{I}_{m-1}$. We remark that the interpretation of the estimable ambiguity parameters does not change as a consequence of the between-satellite differencing, because in the

undifferenced model they are already estimable as between-satellite differences, see Table 21.3 in case of PPP and (21.87) in case of PPP-RTK.

In the multifrequency case, with the ionospheric delays as unknown parameters, the between-satellite differencing applied to the PPP-RTK model in Table 21.5 results in the model presented in Table 21.6.

21.5.2 Double and Triple Differencing

Double-Differenced Relative Positioning Model

For the Relative Positioning models, as discussed in Sect. 21.4, between-satellite differencing can be applied as well, as to remove the receiver-dependent parameters from the models. Alternatively, since multiple receivers are involved that observe the same satellites, one may difference the observations of the same satellite between each receiver and a chosen pivot receiver. This is *between-receiver* differencing (Fig. 21.8b), which

Table 21.6 Full-rank, between-satellite differenced, ionosphere-float, multifrequency, PPP-RTK model

$$E \begin{pmatrix} \mathbf{D}_m^\top \Delta \tilde{\mathbf{p}}_{r,1}(t) \\ \vdots \\ \mathbf{D}_m^\top \Delta \tilde{\mathbf{p}}_{r,f}(t) \\ \hline \mathbf{D}_m^\top \Delta \tilde{\boldsymbol{\varphi}}_{r,1}(t) \\ \vdots \\ \mathbf{D}_m^\top \Delta \tilde{\boldsymbol{\varphi}}_{r,f}(t) \end{pmatrix} = \begin{bmatrix} \mathbf{D}_m^\top \mathbf{G}_r(t) & \mu_1 \mathbf{I}_{m-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}_m^\top \mathbf{G}_r(t) & \mu_f \mathbf{I}_{m-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \hline \mathbf{D}_m^\top \mathbf{G}_r(t) & -\mu_1 \mathbf{I}_{m-1} & \lambda_1 \mathbf{I}_{m-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}_m^\top \mathbf{G}_r(t) & -\mu_f \mathbf{I}_{m-1} & \mathbf{0} & \cdots & \lambda_f \mathbf{I}_{m-1} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_r(t) \\ \hline \mathbf{D}_m^\top \tilde{\mathbf{I}}_r(t) \\ \hline \tilde{N}_{r,1} \\ \vdots \\ \tilde{N}_{r,f} \end{bmatrix}$$

removes the satellite-dependent parameters from the models. This means that both receiver-dependent and satellite-dependent parameters can be eliminated by either taking the between-receiver difference of two between-satellite differences, or taking the between-satellite difference of two between-receiver (Fig. 21.8c). As a consequence one obtains the well-known *double-differenced* positioning model.

In case the ionosphere is float, the double-differenced model has the same structure as the between-satellite differenced PPP-RTK model in Table 21.6, but instead of single-differenced observables the observables are double differenced. In addition, the ionospheric parameters are estimated as double differences as well (the estimable ambiguities are already double differenced in the full-rank between-satellite differenced model). For a double-differenced model, it is implicitly assumed that the network or baseline is of regional size such that the geometry matrices in the design matrix are identical for all receivers, that is, $\mathbf{G}_1(t) = \cdots = \mathbf{G}_n(t) \doteq \mathbf{G}(t)$, and relative position/ZTD parameters are estimated, that is, $\tilde{\mathbf{x}}_r(t) = \mathbf{x}_r(t) - \mathbf{x}_1(t)$.

Triple-Differenced Relative Positioning Model

From the double-differenced model, one could go one step further, by taking differences of two double differences in time, so as to eliminate the ambiguity

parameters from the relative positioning model (provided that no cycle slips have occurred between the two epochs). As a result one obtains the *triple-differenced* model (Fig. 21.8d). However, triple differencing removes the possibility of taking advantage of the integer nature of these double-differenced ambiguities, which is the key requirement for obtaining high-precision positions. Other drawbacks of triple differencing is that it is only possible to estimate the receiver's position *change in time* and that it creates time correlation between the observations. Therefore it not recommended to base the relative positioning model on triple-differenced observations.

21.5.3 Redundancy of the Differenced Models

The *redundancy* (number of observations minus number of estimable parameters) of the differenced models is exactly identical to those of the undifferenced versions of the models, as the models are reduced with the same number of estimable parameters as observations. This, however, only applies to the *single-epoch* case. In the multiepoch case the redundancy of the undifferenced models start to outperform that of the differenced models if temporal constraints on the parameters are included [21.69].

21.6 The Positioning Concepts Related

This chapter provided an overview of the models underlying the various positioning concepts. We end this chapter with a summary of how these positioning concepts are related to each other. Figure 21.9 presents in a schematic way the various positioning concepts. They basically differ in the way whether they are dominated by either code data or phase data, and at the scale the correction data are provided by the reference stations.

21.6.1 Global Positioning: SPP/PPP

At a *global* scale, *absolute* positioning can be realized by means of SPP or PPP. In case of SPP, based on code observations, global reference network data are employed in the form of the broadcast navigation data. In other words, the orbits, clocks, and atmospheric data that are broadcast by the GNSS satellites are products that are determined by the GNSS ground control network. PPP employs phase observations in addition to code, as well as precise corrections for orbit, clocks,

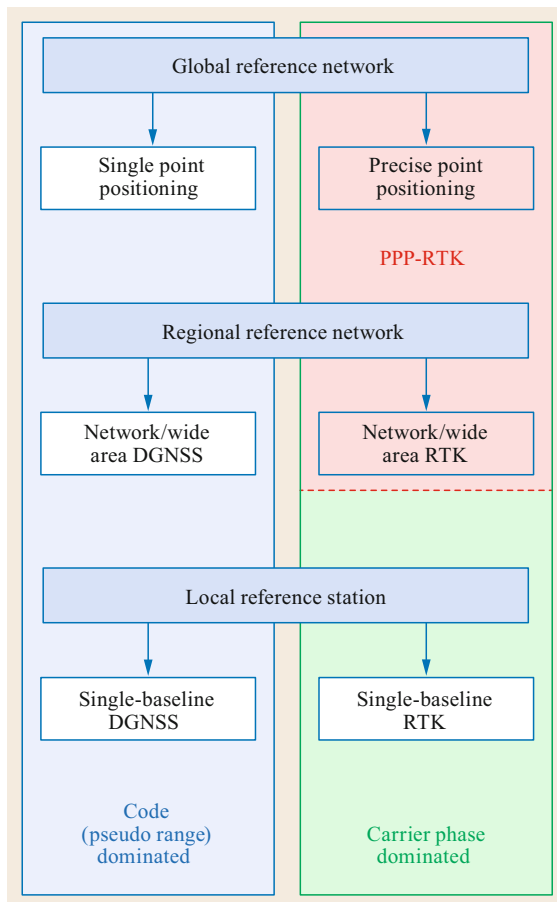


Fig. 21.9 GNSS-based positioning concepts: SPP versus PPP that are both based on products provided by global reference network data (*top*); network DGNSS versus network RTK that are both based on products provided by regional reference network data (*middle*); single-baseline DGNSS versus single-baseline RTK that are both based on local reference station data (*bottom*). PPP-RTK can be regarded as a method that is conceptually equivalent to PPP, but provides the positioning accuracy of network/wide-area RTK

and ionosphere, which are also products of a global reference network (e.g., the IGS network).

21.6.2 Regional Positioning: Network DGNSS/RTK

At a more *regional* scale, covering an area with a radius of typically 500 km or less, we have the code-dominated network DGNSS technique, versus the phase-dominated network RTK technique, where positioning is done relative to a network of reference stations [21.75]. Correction data are determined by the network and

transmitted to users operating within the coverage area of the network. Network DGNSS is also known as wide-area DGNSS [21.76]. In case of network RTK the network processing is based on ambiguity resolution (Chap. 23) as to provide the most precise corrections to users, who employ ambiguity resolution themselves as to obtain positions with centimeter level accuracy. In practice several network-RTK implementations exist (Chap. 26). Crucial to the performance of both network DGNSS and network RTK are the (quality of the) corrections for the differential ionospheric delays, which are determined by the network over the coverage area.

21.6.3 Local Positioning: Single-Baseline DGNSS/RTK

At a *local* scale, the reference data are provided by a single reference station, located in the vicinity of the user receiver, such that the differential ionospheric delays can be neglected. This leads to code-dominated single-baseline DGNSS, and phase-dominated single-baseline RTK, the latter method relying on phase integer ambiguity resolution. RTK positioning is also referred to as *carrier-phase based DGNSS* for which the differential ionospheric delays may be neglected for baseline lengths up to about 10 km (under average ionospheric conditions). The maximum baseline distance of single-baseline DGNSS, which is at most about 100 km, is longer than for single-baseline RTK, since the noisier code data allow more residual differential ionospheric delays than the precise phase data in case of RTK. Extending the operational distance of single-baseline RTK is possible, up to hundreds of kilometers [21.55], however then the ionospheric delays need to be modeled as unknown parameters, leading to longer convergence times.

21.6.4 Global/Regional Positioning: PPP-RTK

PPP-RTK can be considered as a mixture of PPP and RTK: it is conceptually PPP, but based on resolving the integer phase ambiguities as to obtain the positioning accuracy of (network) RTK. PPP-RTK can either be based on global or regional network products, where satellite phase bias corrections are crucial for integer ambiguity resolution and therefore centimeter level accuracy. In the absence of these satellite phase biases, PPP-RTK reduces to standard PPP. In a triple- or higher frequency case also satellite code bias corrections are required. Essential to fast integer ambiguity resolution are precise ionospheric corrections. In contrast to network RTK, in case of PPP-RTK the correction infor-

Table 21.7 Typical values for GPS positioning accuracy, convergence time, and coverage area for different positioning concepts. Note: SF = single-frequency; DF = dual-frequency

Positioning concept	Accuracy (1-sigma)	Convergence time	Coverage area
SF SPP	< 10 m	Instantaneous	Global
SF PPP (GIM-based)	1–2 dm	< 10 min	Global
DF PPP (ionosphere-float)	< 1 dm	30 min (static) 60 min (kin.)	Global
Single-baseline (code-based) DGNSS	1–5 m	Instantaneous	Regional/local
Wide area DGNSS	0.5–2 m	Instantaneous	Regional
SF RTK-short baseline	< 1 dm	10 min	Local
DF RTK-short baseline	< 1 dm	Instantaneous to few min	Local
Network RTK	< 1 dm	< 10 min	Regional
SF PPP-RTK (precise iono corrections)	< 1 dm	< 10 min	Regional
DF PPP-RTK	< 1 dm	30 min (static)	Global
(ionosphere-float)	< 1 dm	90 min (kin.)	Regional

mation is provided to the user in the parameter or state space, whereas in case of network RTK this is in the observation space [21.57]. A drawback of transmitting the corrections in the observation space is that a higher update rate is required than for corrections in the parameter space, as some parameters are (almost) time constant.

21.6.5 Accuracy of the Positioning Concepts

The attainable *positioning accuracy* using the concepts in the left column of Fig. 21.9 is driven by the precision of the pseudorange data, while the accuracy of the concepts in the right column is driven by the carrier-phase precision. Table 21.7 summarizes typical values of the accuracy, which can be obtained using the discussed positioning concepts. These numbers hold after a certain (convergence) time that is needed for the position to attain a certain accuracy. The positioning accuracy of single-frequency (GPS) PPP incorporating global iono-

spheric corrections is typically at decimeter level within a few minutes [21.77]. The positioning accuracy of dual-frequency (GPS) PPP can reach centimeter level however only after a convergence time that may last for more than 30 min [21.52]. Decimeter-level accuracy of dual-frequency (GPS) PPP after a convergence of 40 min was demonstrated for a kinematic receiver in [21.78]. A similar level of accuracy of dual-frequency PPP based on convergence times of 10–30 min was demonstrated by [21.79]. The positioning accuracy of single-baseline DGNSS (DGPS) is 1–5 m and can be obtained instantaneously. With network or wide-area DGPS the accuracy lies in the range 0.5–2 m [21.80]. In case of dual-frequency integer ambiguity resolution enabled PPP centimeter-level accuracy is feasible after about 30 min in case of a static receiver, and about 90 min in case of a kinematic receiver [21.81]. The convergence time of tens of minutes in case of dual-frequency PPP and PPP-RTK is due to the presence of the ambiguities, next to the ionospheric delays.

References

- 21.1 C. Hegarty, E. Powers, B. Fonville: Accounting for timing biases between GPS, modernized GPS, and Galileo signals, Proc. 36th Annu. PTTI Meet., Washington DC (2004) pp. 307–317
- 21.2 R.E. Phelts, G.X. Gao, G. Wong, L. Heng, T. Walter, P. Enge, S. Erker, S. Thoelet, F. Meurer: New GPS signals – Aviation grade chips of the Block IIF, Inside GNSS 5(5), 36–45 (2010)
- 21.3 R. Piriz, M. Cueto, V. Fernandez, P. Tavella, I. Sesia, G. Cerretto, J. Hahn: GPS/Galileo interoperability: GGT0, timing biases and GIOVE-A experience, Proc. 38th Annu. Precise Time Time Interval (PTTI) Meet., Washington DC (2007) pp. 49–68
- 21.4 R. Dach, S. Schaer, S. Lutz, M. Meindl, G. Beutler: Combining the observations from different GNSS, Proc. EUREF 2010 Symp., Gävle (2010)
- 21.5 D. Odijk, P.J.G. Teunissen, L. Huisman: First results of mixed GPS+GIOVE single-frequency RTK in Australia, J. Spatial Sci. 57(1), 3–18 (2012)
- 21.6 S. Schaer: Mapping and Predicting the Earth's Ionosphere Using the Global Positioning System, Ph.D. Thesis (Astronomical Institute, Univ. Berne, Berne, Switzerland 1999)
- 21.7 O. Montenbruck, A. Hauschild: Code biases in multi-GNSS point positioning, Proc. ION ITM 2013, San Diego (ION, Virginia 2013) pp. 616–628

- 21.8 J. Hahn, E.D. Powers: Implementation of the GPS to Galileo time offset (GGTO), Proc. IEEE Int. FCS PTI Syst. Appl. Meet., Vancouver (2005) pp. 33–37
- 21.9 P. Defraigne, W. Aerts, G. Cerretto, G. Signorile, E. Cantoni, I. Sesia, P. Tavella, A. Cernigliaro, A. Samperi, J.M. Sleewaegen: Advances on the use of Galileo signals in time metrology: Calibrated time transfer and estimation of UTC and GGTO using a combined commercial GPS–Galileo receiver, Proc. 45th Annu. PTI Syst. Appl. Meet., Bellevue (2013) pp. 256–262
- 21.10 M. Aoki: QZSS: The Japanese quasi-zenith satellite system – Program updates and current status, Proc. 5th Meet. Int. Comm. GNSS (ICG), Torino (UNOOSA, Vienna 2010)
- 21.11 K.V.D. Rajarajan, S.C.N.T. Rathnakara, A.S. Ganeshan: Modeling of IRNSS system time-offset with respect to other GNSS, Contr. Theory Inf. **5**(2), 10–17 (2015)
- 21.12 B. Hofmann-Wellenhof, H. Lichtenegger, E. Wasle: *GNSS – Global Navigation Satellite Systems, GPS, GLONASS, Galileo and More* (Springer, Wien 2008)
- 21.13 O. Montenbruck, P. Steigenberger: The BeiDou navigation message, J. Glob. Position. Syst. **12**(1), 1–12 (2013)
- 21.14 A.E. Zinoviev: Using GLONASS in combined GNSS receivers: Current status, Proc. ION GNSS 2005, Long Beach (ION, Virginia 2005) pp. 1046–1057
- 21.15 G. Gendt, Z. Altamimi, R. Dach, W. Söhne, T. Springer: GGSP: Realisation and maintenance of the Galileo terrestrial reference frame, Adv. Space Res. **47**(2), 174–185 (2011)
- 21.16 Indian Regional Navigation Satellite System – Signal in Space ICD for Standard Positioning Service, (Indian space research organization, Bangalore, 2014)
- 21.17 V. Vdovin: National reference systems of the Russian federation, used in GLONASS, including the user and fundamental segments, Proc. 8th Meet. Int. Comm. GNSS (ICG), Working Group D, Dubai (UNOOSA, Vienna 2013) pp. 1–11
- 21.18 P.J. de Jonge: A Processing Strategy for the Application of the GPS in Networks, Ph.D. Thesis (Netherlands Geodetic Commission, Delft 1998), Publications on Geodesy, 46
- 21.19 A. Leick, L. Rapoport, D. Tatarnikov: *GPS Satellite Surveying*, 4th edn. (John Wiley, Hoboken 2015)
- 21.20 P.J. Buist: Multi-Platform Integrated Positioning and Attitude Determination Using GNSS, Ph.D. Thesis (Delft University of Technology, Delft 2013)
- 21.21 T. Ebinuma: Precision Spacecraft Rendezvous Using Global Positioning System: An Integrated Hardware Approach, Ph.D. Thesis (University of Texas, Austin 2001)
- 21.22 Navstar GPS Space Segment/Navigation User Interfaces, Interface Specification IS-GPS-200H (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo 2013)
- 21.23 European Global Navigation Satellite Systems Agency: European GNSS (Galileo) Open Service Signal in Space Interface Control Document, OS SIS ICD, Iss. 1. (2010)
- 21.24 O. Montenbruck, P. Steigenberger, S. Riley: IRNSS orbit determination and broadcast ephemeris assessment, Proc. ION ITM 2015, Dana Point (2015) pp. 185–193
- 21.25 J. Ray, K. Senior: Geodetic techniques for time and frequency comparisons using GPS phase and code measurements, Metrologia **42**(4), 215–232 (2005)
- 21.26 A.Q. Le: Achieving decimetre accuracy with single frequency standalone GPS positioning, Proc. ION GNSS 2004, Long Beach (ION, Virginia 2004) pp. 1881–1892
- 21.27 E.D. Kaplan, C.J. Hegarty: *Understanding GPS: Principles and Applications*, 2nd edn. (Artech House, Boston, London 2006)
- 21.28 A. Tetewsky, J. Ross, A. Soltz, N. Vaughn, J. Anszperger, C. O'Brien, D. Graham, D. Craig, J. Lozow: Making sense of inter-signal corrections, Inside GNSS **4**(4), 37–48a (2009)
- 21.29 BeiDou Navigation Satellite System Signal In Space Interface Control Document – Open Service Signal, v.2.0 (China Satellite Navigation Office, 2013)
- 21.30 Global Navigation Satellite System GLONASS–Interface Control Document, Vol. 5.1 (Russian Institute of Space Device Engineering, Moscow, 2008)
- 21.31 U. Rossbach: Positioning and Navigation Using the Russian Satellite System GLONASS, Ph.D. Thesis (Universität der Bundeswehr München, Munich 2000)
- 21.32 C.H. Yinger, W.A. Feess, R.D. Esposti, A. Chasko, B. Cosentino, D. Syse, B. Wilson, B. Wheaton: GPS satellite interfrequency biases, Proc. ION AM 1999, Cambridge (ION, Virginia 1999) pp. 347–354
- 21.33 O. Montenbruck, A. Hauschild, P. Steigenberger: Differential code bias estimation using multi-GNSS observations and global ionosphere maps, Navigation **61**(3), 191–201 (2014)
- 21.34 E. Sardón, A. Rius, N. Zarraoa: Estimation of the transmitter and receiver differential biases and the ionospheric total electron content from global positioning system observations, Radio Sci. **29**(3), 577–586 (1994)
- 21.35 M. Ge, G. Gendt, M. Rothacher, C. Shi, J. Liu: Resolution of GPS carrier-phase ambiguities in precise point positioning (PPP) with daily observations, J. Geod. **82**(7), 389–399 (2008)
- 21.36 J. Kouba, P. Héroux: Precise point positioning using IGS orbit and clock products, GPS Solutions **5**(2), 12–28 (2001)
- 21.37 A.J. Mannucci, B.D. Wilson, C.D. Edwards: A new method for monitoring the Earth's ionospheric total electron content using the GPS global network, Proc. ION GPS 1993, Salt Lake City (ION, Virginia 1993) pp. 1323–1332
- 21.38 J.A. Klobuchar: Ionospheric time-delay algorithm for single-frequency GPS users, IEEE Trans. Aerosp. Electron. Syst. **23**(3), 325–331 (1987)
- 21.39 A. Angrisano, S. Gaglione, C. Gioia, M. Massaro, U. Robustelli: Assessment of NeQuick ionospheric model for Galileo single-frequency users, Acta Geophysica **61**(6), 1457–1476 (2013)

- 21.40 J. Saastamoinen: Atmospheric correction for the troposphere and stratosphere in radio ranging of satellites. In: *The Use of Artificial Satellites for Geodesy*, AGU Geophys. Monogr., Vol. 15, ed. by H.W. Henriksen, A. Mancini, B.M. Chovitz (The American Geophysical Union, Washington 1972) pp. 247–251
- 21.41 S. Bancroft: An algebraic solution of the GPS equations, *IEEE Trans. Aerosp. Electron. Syst.* **21**(7), 56–59 (1985)
- 21.42 L.O. Krause: A direct solution to GPS-type navigation equations, *IEEE Trans. Aerosp. Electron. Syst.* **23**(2), 225–232 (1987)
- 21.43 A. Kleusberg: Analytical GPS navigation solution, Quo vadis geodesia...? In: *Festschrift for Erik W. Grafarend on the Occasion of his 60th Birthday*, ed. by F. Krumm, V.S. Schwarze (Univ. Stuttgart, Stuttgart 1999) pp. 247–251
- 21.44 D. Odijk, P.J.G. Teunissen: Characterization of between-receiver GPS-Galileo inter-system biases and their effect on mixed ambiguity resolution, *GPS Solutions* **17**(4), 521–533 (2013)
- 21.45 R.B. Langley: Dilution of precision, *GPS World* **10**(5), 52–59 (1999)
- 21.46 P.J.G. Teunissen: A proof of Nielsen's conjecture on the GPS dilution of precision, *IEEE Trans. Aerosp. Electron. Syst.* **34**(2), 693–695 (1998)
- 21.47 P.J.G. Teunissen: GPS op afstand bekeken (in Dutch). In: *Een halve eeuw in de goede richting – Lustrumboek Snellius 1985–1990*, (DUM, Delft 1990) pp. 215–233
- 21.48 P. Héroux, J. Kouba: GPS precise point positioning with a difference, *Proc. Geomatics'95*, Ottawa (1995) pp. 1–11
- 21.49 J.F. Zumberge, M.B. Hefflin, D.C. Jefferson, M.M. Watkins, F.H. Webb: Precise point positioning for the efficient and robust analysis of GPS data from large networks, *J. Geophys. Res.* **102**(B3), 5005–5017 (1997)
- 21.50 A.E. Niell: Global mapping functions for the atmosphere delay at radio wavelengths, *J. Geophys. Res.* **101**(B2), 3227–3246 (1996)
- 21.51 L. Wanninger: Carrier-phase inter-frequency biases of GLONASS receivers, *J. Geod.* **86**(2), 139–148 (2012)
- 21.52 H. van der Marel, P.F. de Bakker: Single-vs. dual-frequency precise point positioning – What are the tradeoffs between using L1-only and L1+L2 for PPP?, *GNSS Solutions* **7**(4), 30–35 (2012)
- 21.53 D. Odijk, P.J.G. Teunissen, A. Khodabandeh: Galileo IOV RTK positioning: Standalone and combined with GPS, *Survey Rev.* **46**(337), 267–277 (2014)
- 21.54 R. Odolinski, P.J.G. Teunissen, D. Odijk: Combined GPS and BeiDou instantaneous RTK positioning, *Navigation* **61**(2), 135–148 (2014)
- 21.55 T. Takasu, A. Yasuda: Kalman-filter-based integer ambiguity resolution strategy for long-baseline RTK with ionosphere and troposphere estimation, *Proc. ION GNSS 2010*, Portland (ION, Virginia 2010) pp. 161–171
- 21.56 R. Odolinski, P.J.G. Teunissen, D. Odijk: Combined GPS+BDS+Galileo+QZSS for long baseline RTK positioning, *Proc. ION GNSS 2014*, Tampa (ION, Virginia 2014) pp. 2326–2340
- 21.57 G. Wübbena, M. Schmitz, A. Bagge: PPP-RTK: Precise point positioning using state-space representation in RTK networks, *Proc. ION GNSS 2005*, Long Beach (ION, Virginia 2005) pp. 2584–2594
- 21.58 O. Montenbruck, P. Steigenberger, A. Hauschild: Broadcast versus precise ephemerides: A multi-GNSS perspective, *GPS Solutions* **19**(2), 321–333 (2015)
- 21.59 P.J.G. Teunissen, A. Kleusberg (Eds.): *GPS for Geodesy*, 2nd edn. (Springer, Berlin 1998)
- 21.60 P.J.G. Teunissen: The geometry-free GPS ambiguity search space with a weighted ionosphere, *J. Geod.* **71**(6), 370–383 (1997)
- 21.61 W. Lindlohr, D. Wells: GPS design using undifferenced carrier beat phase observations, *Manuscripta Geodaetica* **10**(4), 255–295 (1985)
- 21.62 C.C. Goad: Precise relative position determination using global positioning system carrier phase measurements in a nondifference mode, *Proc. 1st Int. Symp. Precise Position. Glob. Position. Syst.*, Rockville, ed. by C. Goad (U.S. Department of Commerce, Maryland 1985) pp. 347–356
- 21.63 G. Blewitt: Carrier phase ambiguity resolution for the global positioning system applied to geodetic baselines up to 2000 km, *J. Geophys. Res.* **94**(B8), 10187–10203 (1989)
- 21.64 P.J.G. Teunissen: The least-squares ambiguity decorrelation adjustment: A method for fast GPS integer ambiguity estimation, *J. Geod.* **70**(1/2), 65–82 (1995)
- 21.65 P.J.G. Teunissen, D. Odijk, B. Zhang: PPP-RTK: Results of CORS network-based PPP with integer ambiguity resolution, *J. Aeronaut., Astronaut. Aviat., Ser. A* **42**(4), 223–230 (2010)
- 21.66 B. Zhang, P.J.G. Teunissen, D. Odijk: A novel un-differenced PPP-RTK concept, *RIN J. Navig.* **64**(Supplement S1), S180–S191 (2011)
- 21.67 D. Odijk, P.J.G. Teunissen, B. Zhang: Single-frequency integer ambiguity resolution enabled GPS precise point positioning, *J. Surv. Eng.* **138**(4), 193–202 (2012)
- 21.68 C. Rocken, R. Ware, T. van Hove, F. Solheim, C. Alber, J. Johnson: Sensing atmospheric water vapor with the global positioning system, *Geophys. Res. Lett.* **20**(23), 2631–2634 (1993)
- 21.69 D. Odijk, B. Zhang, A. Khodabandeh, R. Odolinski, P.J.G. Teunissen: On the estimability of parameters in undifferenced GNSS network and PPP-RTK user models by means of S-system theory, *J. Geod.* **90**(1), 15–44 (2016)
- 21.70 P. Collins, S. Bisnath, F. Lahaye, P. Héroux: Undifferenced GPS ambiguity resolution using the decoupled clock model and ambiguity datum fixing, *Navigation* **57**(2), 123–135 (2010)
- 21.71 J. Geng, F.N. Teferle, X. Meng, A.H. Dodson: Towards PPP-RTK: Ambiguity resolution in real-time precise point positioning, *Adv. Space Res.* **47**(10), 1664–1673 (2011)
- 21.72 S. Loyer, F. Perosanz, F. Mercier, H. Capdeville, J.-C. Marty: Zero-difference GPS ambiguity resolution at CNES-CLS IGS Analysis Center, *J. Geod.* **86**(11), 991–

- 1003 (2012)
- 21.73 P.J.G. Teunissen, A. Khodabandeh: Review and principles of PPP-RTK methods, *J. Geod.* **89**(3), 217–240 (2015)
- 21.74 R. Christensen: *Linear Models for Multivariate, Time Series, and Spatial Data* (Springer, Berlin 1991)
- 21.75 H. Landau, X. Chen, S. Klose, R. Leandro, U. Volath: Trimble's RTK and DGPS solutions in comparison with precise point positioning, *Observing our Changing Earth*, Proc. Int. Assoc. Geod. Symp. 133, Perugia, ed. by M.G. Sideris (Springer, Berlin 2009) pp. 709–718
- 21.76 C. Kee, B.W. Parkinson: Wide area differential GPS (WADGPS): Future navigation system, *IEEE Trans. Aerosp. Electron. Syst.* **32**(2), 795–808 (1996)
- 21.77 R.J.P. van Bree, C.C.J.M. Tiberius: Real-time single-frequency precise point positioning: Accuracy assessment, *GPS Solutions* **16**(2), 259–266 (2012)
- 21.78 M.O. Kechine, C.C.J.M. Tiberius, H. van der Marel: Experimental verification of internet-based global differential GPS, Proc. ION GPS 2003, Portland (ION, Virginia 2003) pp. 28–37
- 21.79 D. Lapucha, K. de Jong, X. Liu, T. Melgard, O. Oerpen, E. Vigen: Recent advances in wide area real-time precise positioning, *TransNav Int. J. Marine Navig. Safety Sea Transp.* **5**(1), 87–92 (2011)
- 21.80 J.D. Bossler, J.R. Jensen, R.B. McMaster, C. Rizos (Eds.): *Manual of Geospatial Science and Technology* (Taylor Francis, London 2002)
- 21.81 D. Laurichesse: Phase biases estimation for integer ambiguity resolution, Proc. PPP-RTK Open Stand. Symp., Frankfurt am Main (BKG, Frankfurt 2013)

22. Least-Squares Estimation and Kalman Filtering

Sandra Verhagen, Peter J.G. Teunissen

This chapter presents the estimation and filtering principles as used in global navigation satellite system (GNSS) data processing. Estimation and filtering are concerned with retrieving or recovering parameters of interest from noisy measurements. The least-squares (LS) principle is the standard approach for estimating unknown parameters from uncertain data. Various forms of LS estimation, such as partitioned-LS, recursive-LS, constrained-LS, and nonlinear-LS, are discussed.

The parameters of interest, as well as the dominant error sources, are often time varying. If these time variations can be modeled, the parameters can be resolved based on minimum mean squared error prediction, filtering, and smoothing techniques. Of the various such techniques, the Kalman filter is most prominent. It recursively estimates the state of a dynamic system. Different forms of the Kalman filter are discussed, together with its linkage to recursive smoothing techniques. Several GNSS examples are included in support of the general introduction on the principles and properties of LS estimation and Kalman filtering.

22.1	Linear Least-Squares Estimation	639
22.1.1	Least-Squares Principle.....	639
22.1.2	Weighted Least-Squares.....	640
22.1.3	Computation of LS Solution.....	640
22.1.4	Statistical Properties.....	641
22.2	Optimal Estimation	641
22.2.1	Best Linear Unbiased Estimation.....	641
22.2.2	Maximum Likelihood Estimation.....	642
22.2.3	Confidence Regions.....	642
22.3	Special Forms of Least Squares	644
22.3.1	Recursive Estimation.....	644
22.3.2	Estimation with Partitioned Parameter Vector.....	646
22.3.3	Block Estimation.....	647
22.3.4	Constrained Least-Squares.....	647
22.3.5	Rank-Defect Least Squares.....	647
22.3.6	Non-Linear Least-Squares.....	648
22.4	Prediction and Filtering	650
22.4.1	Prediction Problem.....	650
22.4.2	Minimum Mean Squared Error Prediction.....	651
22.4.3	Properties of MMSE Prediction.....	653
22.5	Kalman Filtering	653
22.5.1	Model Assumptions.....	653
22.5.2	The Kalman Filter Recursion.....	654
22.5.3	Kalman Filter Information Form.....	655
22.5.4	Extended Kalman Filter.....	656
22.5.5	Smoothing.....	657
	References	659

22.1 Linear Least-Squares Estimation

GNSS observations are, like all empirical data, subject to uncertainty – measurements are never perfect. Moreover, we generally have *redundant* measurements: There are more observations available than strictly needed for estimating the parameters of interest. In this section, we introduce the principle of least-squares (LS) estimation for solving overdetermined systems, that is, estimation problems with redundant measurements. Due to the uncertainty of the measurements, the redundancy generally leads to *inconsistent* systems

of equations; estimating the parameters with different subsets of the observations will then lead to different solutions. The LS method ensures that still a unique solution can be obtained by imposing additional criteria. This section explains the LS principle and its properties.

22.1.1 Least-Squares Principle

The objective is to obtain estimates of n unknown parameters x_j , $j = 1, \dots, n$ from a set of m measurements

$y_i, i = 1, \dots, m$. If there is a linear relationship between the measurements and the unknown parameters, the following linear system of equations is obtained

$$\mathbf{y} = \mathbf{A}\mathbf{x} \quad (22.1)$$

with the m -vector $\mathbf{y} = [y_1, \dots, y_m]^\top$, unknown parameter vector $\mathbf{x} = [x_1, \dots, x_n]^\top$, and an $m \times n$ coefficient matrix \mathbf{A} .

In general, the system of equations is overdetermined and

$$m > n = \text{rank}(\mathbf{A}). \quad (22.2)$$

Hence, we assume that the matrix \mathbf{A} is of full column rank.

Due to measurement errors, this generally implies that the system in (22.1) is inconsistent, since it is not possible to find a solution \mathbf{x} which exactly reproduces \mathbf{y} , and consequently we can only have $\mathbf{y} \approx \mathbf{A}\mathbf{x}$. The LS principle can then be applied to solve this problem. The idea is to explicitly add the measurement error vector \mathbf{e} to the system of equations

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}. \quad (22.3)$$

Since the measurement errors are unknown, this leads to a total of $m + n$ unknown parameters in m equations and consequently an infinite number of solutions exists for \mathbf{x} . The solution of \mathbf{x} is selected using the sum of squares of the entries of \mathbf{e} as a measure of the discrepancy between \mathbf{y} and $\mathbf{A}\mathbf{x}$. More specifically, the LS solution $\hat{\mathbf{x}}$ of \mathbf{x} , which minimizes this sum of squares $\mathbf{e}^\top \mathbf{e} = (\mathbf{y} - \mathbf{A}\mathbf{x})^\top (\mathbf{y} - \mathbf{A}\mathbf{x})$ is chosen [22.1, 2]

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \min_{\mathbf{x} \in \mathbb{R}^n} (\mathbf{y} - \mathbf{A}\mathbf{x})^\top (\mathbf{y} - \mathbf{A}\mathbf{x}) \\ &= (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y}. \end{aligned} \quad (22.4)$$

The difference between \mathbf{y} and adjusted $\hat{\mathbf{y}} = \mathbf{A}\hat{\mathbf{x}}$ is called the vector of LS residuals

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}}. \quad (22.5)$$

In this way, the scalar inconsistency of the LS is measured by $\|\hat{\mathbf{e}}\|^2 = \hat{\mathbf{e}}^\top \hat{\mathbf{e}}$.

It is also possible to give a geometric interpretation to the LS principle. The estimated $\hat{\mathbf{y}}$ with parameter vector $\hat{\mathbf{x}}$ follows as

$$\hat{\mathbf{y}} = \mathbf{A}\hat{\mathbf{x}} = \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} = \mathbf{P}_\mathbf{A} \mathbf{y} \quad (22.6)$$

with $\mathbf{P}_\mathbf{A}$ being an orthogonal projector. This shows that $\hat{\mathbf{y}} = \mathbf{A}\hat{\mathbf{x}}$ is the orthogonal projection of \mathbf{y} onto the range space of \mathbf{A} .

This can also be illustrated as follows. For the observation vector, we have that $\mathbf{y} \in \mathbb{R}^m$, but according to our model, $\mathbf{A}\mathbf{x}$ must be in the range space $R(\mathbf{A})$ spanned by the columns of \mathbf{A} . Therefore, for the solution $\hat{\mathbf{x}}$ it is required that $\mathbf{A}\hat{\mathbf{x}} \in R(\mathbf{A})$, and with LS the solution with the shortest distance to \mathbf{y} is selected. This implies that it is the orthogonal projection of \mathbf{y} onto $R(\mathbf{A})$, as illustrated in Fig. 22.1.

22.1.2 Weighted Least-Squares

In practice it might be that not all observations y_i have the same precision, and then it makes sense to give more weight to observations with higher precision. The principle of weighted least-squares (WLS) allows us to consider different weights by minimizing the weighted sum of squared residuals $\|\mathbf{e}\|_{\mathbf{W}^{-1}}^2 = \mathbf{e}^\top \mathbf{W} \mathbf{e}$, where \mathbf{W} is the weight matrix with the weights $w_{ii} > 0, i = 1, \dots, m$ as diagonal elements. The weight matrix must be positive definite, but is not necessarily a diagonal matrix.

The WLS solution equals

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \min_{\mathbf{x} \in \mathbb{R}^n} (\mathbf{y} - \mathbf{A}\mathbf{x})^\top \mathbf{W} (\mathbf{y} - \mathbf{A}\mathbf{x}) \\ &= (\mathbf{A}^\top \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{W} \mathbf{y}. \end{aligned} \quad (22.7)$$

A logical choice for the weight matrix is $\mathbf{W} = \mathbf{Q}_{yy}^{-1}$ with \mathbf{Q}_{yy} being the variance–covariance matrix of the observables. The reason is that it makes sense to give a larger weight to more precise measurements, and a smaller weight to less precise measurements.

22.1.3 Computation of LS Solution

The WLS problem is to solve the system of normal equations

$$\mathbf{A}^\top \mathbf{W} \mathbf{A} \hat{\mathbf{x}} = \mathbf{A}^\top \mathbf{W} \mathbf{y} \quad (22.8)$$

with normal matrix $\mathbf{N} = \mathbf{A}^\top \mathbf{W} \mathbf{A}$. The right-hand side will be denoted as $\mathbf{r} = \mathbf{A}^\top \mathbf{W} \mathbf{y}$. The WLS solution can be computed by inverting the normal equations

$$\hat{\mathbf{x}} = \mathbf{N}^{-1} \mathbf{r}. \quad (22.9)$$

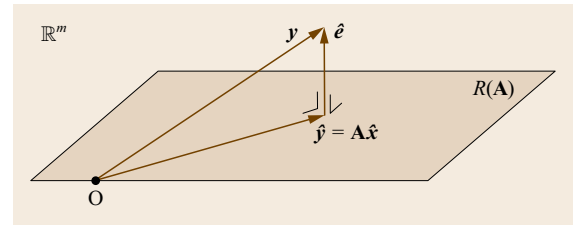


Fig. 22.1 Geometry of least-squares: $\mathbf{y} = \mathbf{A}\hat{\mathbf{x}} + \hat{\mathbf{e}}$ (after [22.2])

This, however, is not advisable from a numerical point of view, since the explicit computation of the inverse of the normal matrix can be time-consuming and prone to round-off errors. A better alternative is to work with the *Cholesky* decomposition of the positive-definite normal matrix \mathbf{N} , $\mathbf{N} = \mathbf{G}\mathbf{G}^\top$, [22.3].

Solving the normal equations with the Cholesky decomposition starts with splitting the original system into two systems

$$\mathbf{G}\hat{\mathbf{g}} = \mathbf{r} \quad \text{and} \quad \mathbf{G}^\top \hat{\mathbf{x}} = \hat{\mathbf{g}}, \quad (22.10)$$

where \mathbf{G} is the lower triangular matrix or Cholesky factor. First, forward substitution is applied to compute the entries of $\hat{\mathbf{g}}$ starting with the first entry $\hat{g}_1 = r_1/G_{11}$, then $\hat{g}_2 = (r_2 - G_{21}\hat{g}_1)/G_{22}$, etc. (where G_{ij} is entry (i,j) from matrix \mathbf{G}). Once $\hat{\mathbf{g}}$ has been computed, backward substitution is applied to compute the entries of $\hat{\mathbf{x}}$ in a similar fashion using the right-hand side of (22.10), but then starting with the last entry since \mathbf{G}^\top is an upper triangular matrix.

Alternative methods for computing the LS solution make use of different matrix decompositions, such as the orthogonal QR factorization, $\mathbf{A} = \mathbf{Q}\mathbf{R}$, or the singular value decomposition (SVD), $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. For a discussion on their principles and properties, see [22.3–5].

22.1.4 Statistical Properties

In order to be able to solve the inconsistent linear system of equations $\mathbf{y} = \mathbf{A}\mathbf{x}$, the measurement error vector \mathbf{e} was introduced. The WLS solution is obtained by minimizing the weighted squared norm $\mathbf{e}^\top \mathbf{W}\mathbf{e}$. In general, it is assumed that the measurement errors are random and zero on average. In other words, if the

measurements were to be repeated many times under similar circumstances, the mean errors would become zero. The mean or mathematical expectation $E(\cdot)$ of \mathbf{e} is thus assumed to be zero, $E(\mathbf{e}) = \mathbf{0}$. Consequently, the mean of \mathbf{y} becomes

$$E(\mathbf{y}) = E(\mathbf{A}\mathbf{x} + \mathbf{e}) = \mathbf{A}\mathbf{x}, \quad (22.11)$$

since \mathbf{x} is the deterministic vector with unknown parameters. The means of the WLS estimators follow as

$$E(\hat{\mathbf{x}}) = (\mathbf{A}^\top \mathbf{W}\mathbf{A})^{-1} \mathbf{A}^\top \mathbf{W}E(\mathbf{y}) = \mathbf{x}, \quad (22.12)$$

$$E(\hat{\mathbf{y}}) = \mathbf{A}E(\hat{\mathbf{x}}) = \mathbf{A}\mathbf{x}, \quad (22.13)$$

$$E(\hat{\mathbf{e}}) = E(\mathbf{y}) - E(\hat{\mathbf{y}}) = \mathbf{0}. \quad (22.14)$$

This shows that WLS provides *unbiased* estimators.

In the consistent linear system of equations $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, \mathbf{y} and \mathbf{e} are the random vectors whereas \mathbf{x} is deterministic. This implies that $\mathbf{Q}_{yy} = \mathbf{Q}_{ee}$. By application of the variance propagation law [22.2], the variance–covariance matrices of the WLS estimators can be derived as

$$\mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} = (\mathbf{A}^\top \mathbf{W}\mathbf{A})^{-1} \mathbf{A}^\top \mathbf{W}\mathbf{Q}_{yy}\mathbf{W}\mathbf{A}(\mathbf{A}^\top \mathbf{W}\mathbf{A})^{-1}, \quad (22.15)$$

$$\mathbf{Q}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \mathbf{A}\mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\mathbf{A}^\top = \mathbf{P}_\mathbf{A}\mathbf{Q}_{yy}\mathbf{P}_\mathbf{A}^\top, \quad (22.16)$$

$$\mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}} = \mathbf{P}_\mathbf{A}^\perp \mathbf{Q}_{yy} \mathbf{P}_\mathbf{A}^{\perp\top}, \quad (22.17)$$

with orthogonal projector

$$\mathbf{P}_\mathbf{A}^\perp = \mathbf{I}_m - \mathbf{P}_\mathbf{A} \quad \text{and} \quad \mathbf{P}_\mathbf{A} = \mathbf{A}(\mathbf{A}^\top \mathbf{W}\mathbf{A})^{-1} \mathbf{A}^\top \mathbf{W}.$$

The variance–covariance matrices describe the precision of the WLS estimators.

22.2 Optimal Estimation

22.2.1 Best Linear Unbiased Estimation

The principle of *best linear unbiased estimation* (BLUE) is based on the following requirements for the estimator [22.2, 6]. First, the estimator must be *unbiased*, implying that $E(\hat{\mathbf{x}}) = \mathbf{x}$. Furthermore, the condition that the estimator must be *linear* means that the estimator $\hat{\mathbf{x}}$ is a linear function of \mathbf{y} . This can be further generalized by considering the situation where the actual parameters of interest are a linear function of \mathbf{x} , say

$$\mathbf{z} = \mathbf{F}^\top \mathbf{x} + \mathbf{f}. \quad (22.18)$$

The estimator $\hat{\mathbf{z}}$ is called a *linear unbiased* estimator if it is unbiased and a linear function of the observations \mathbf{y} as well

$$E(\hat{\mathbf{z}}) = \mathbf{z} \quad \text{and} \quad \hat{\mathbf{z}} = \mathbf{L}^\top \mathbf{y} + \mathbf{l}. \quad (22.19)$$

The linear unbiased estimator $\hat{\mathbf{z}}$ is called *best* if it has the smallest variance of all linear unbiased estimators. In Sect. 22.1.2, it was already argued that $\mathbf{W} = \mathbf{Q}_{yy}^{-1}$ would be a good choice for the weight matrix. And indeed, it can be proven that the WLS estimator becomes the BLUE when the weight matrix is chosen equal to the inverse variance–covariance matrix of the observables.

Thus, the BLUE of $\mathbf{z} = \mathbf{F}^\top \mathbf{x} + \mathbf{f}$ becomes

$$\hat{\mathbf{z}} = \mathbf{F}^\top \hat{\mathbf{x}} + \mathbf{f} \quad (22.20)$$

with the BLUE of \mathbf{x} given as

$$\hat{\mathbf{x}} = (\mathbf{A}^\top \mathbf{Q}_{yy}^{-1} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{Q}_{yy}^{-1} \mathbf{y}. \quad (22.21)$$

WLS estimation and BLUE are thus identical when $\mathbf{W} = \mathbf{Q}_{yy}^{-1}$. The variance–covariance matrices of $\hat{\mathbf{x}}$ and $\hat{\mathbf{e}}$ from (22.15) and (22.17) become then

$$\mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} = (\mathbf{A}^\top \mathbf{Q}_{yy}^{-1} \mathbf{A})^{-1}, \quad (22.22)$$

$$\mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}} = \mathbf{Q}_{yy} - \mathbf{Q}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}. \quad (22.23)$$

As it is common practice to assume the measurements to be normally distributed, $\mathbf{y} \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \mathbf{Q}_{yy})$, the estimators, being linear functions of \mathbf{y} , will then also be normally distributed

$$\hat{\mathbf{x}} \sim \mathcal{N}(\mathbf{x}, \mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}), \quad (22.24)$$

$$\hat{\mathbf{y}} \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \mathbf{Q}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}), \quad (22.25)$$

$$\hat{\mathbf{e}} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}}). \quad (22.26)$$

Summarizing, BLUE thus has various optimality properties. First, it minimizes the weighted sum of squared residuals. Second, it has the best precision, that is, smallest variance, of all linear unbiased estimators. Moreover, if in addition the model is a Gaussian linear model $\mathbf{y} \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \mathbf{Q}_{yy})$, BLUE will also maximize the likelihood as will be discussed in the next subsection.

22.2.2 Maximum Likelihood Estimation

Another estimation principle is *maximum likelihood estimation* (MLE), which is based on maximizing the

so-called likelihood function of a given observation vector \mathbf{y} . For this purpose, the general structure of the probability density function (PDF) of \mathbf{y} must be known, except for some n unknown parameters \mathbf{x} . The PDF is denoted as $f_y(\cdot|\mathbf{x})$, and it will change depending on the choice for \mathbf{x} . As such, there is a whole family of PDFs, and it is not known to which PDF the observed \mathbf{y} belongs.

The idea is then to select from this family of PDFs, the one which best supports the observed data by considering $f_y(\mathbf{y}|\mathbf{x})$ as function of \mathbf{x} . The likelihood function is now given by a function of \mathbf{x} producing the corresponding probability densities of all PDFs for the same sample \mathbf{y} . The vector \mathbf{x} is then chosen as the one which maximizes the likelihood function

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x} \in \mathbb{R}^n} f_y(\mathbf{y}|\mathbf{x}). \quad (22.27)$$

This is the maximum likelihood estimator of \mathbf{x} . The principle is illustrated in Fig. 22.2.

MLE is a very general estimation principle, which relies on the knowledge of the structure of the PDF (in contrast to WLS and BLUE). In case of the Gaussian linear model, the PDF is given as $f_y(\mathbf{y}|\mathbf{x}) \propto \exp(-\frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_{\mathbf{Q}_{yy}}^2)$. In that case the MLE of \mathbf{x} becomes identical to the BLUE given in (22.21).

22.2.3 Confidence Regions

In Sect. 22.2.1, the variance–covariance matrices of the estimators were presented. These matrices contain all information regarding the precision – the variances on the diagonal – and the correlation between the estimated parameters – by means of the covariances. However, the variance (or its square root, the standard deviation) of

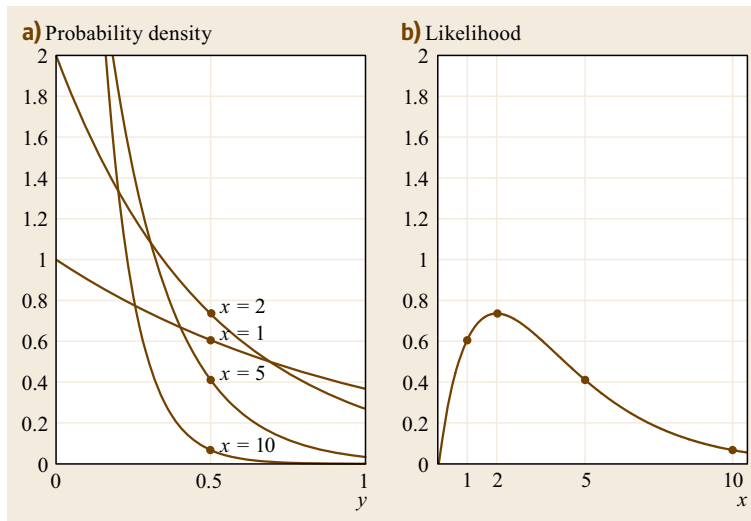


Fig. 22.2a,b PDF and likelihood for exponential distribution. **(a)** Family of PDFs $f_y(\mathbf{y}|\mathbf{x}) = x \exp(-xy)$ with $y \geq 0$, $0 < x < \infty$, for different parameter values x ; **(b)** corresponding likelihood function for $y = 0.5$. The maximum likelihood estimate in this case is equal to $\hat{x} = 2$

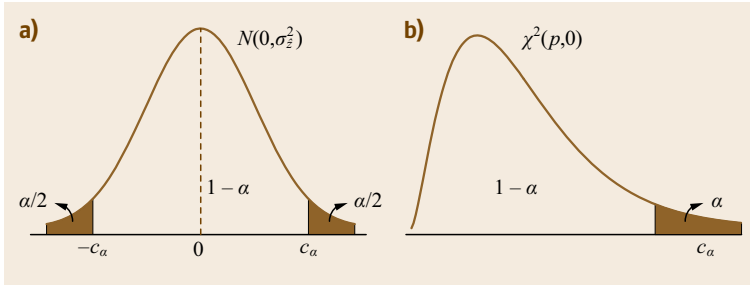


Fig. 22.3 (a) PDF of $N(0, \sigma_z^2)$ with confidence interval $[-c_\alpha, c_\alpha]$; (b) PDF of $\chi^2(p, 0)$ with confidence interval $[0, c_\alpha]$. The confidence level in (a) and (b) is equal to $1 - \alpha$

an estimator does not directly provide information on the probability that the estimation error is smaller than a certain value. Therefore, the concept of *confidence regions* can be used. Starting with the estimation error $\hat{z} - z$ of a single parameter, the $100(1 - \alpha)\%$ confidence interval is defined as

$$CI_\alpha = \{z \in \mathbb{R} \mid -c_\alpha < \hat{z} - z < c_\alpha\} . \quad (22.28)$$

In practice, such a confidence interval is often denoted as $\hat{z} \pm c_\alpha$. It tells you that the estimated value will be within a distance c_α from the true value with a confidence level of $1 - \alpha$, that is,

$$P(|\hat{z} - z| < c_\alpha) = 1 - \alpha . \quad (22.29)$$

With the assumption that the estimator is normally distributed, it is straightforward to determine c_α for a given choice of α , and vice versa. See also Fig. 22.3a. Some common choices are

$$\begin{cases} c_{0.05} &= 1.96\sigma_{\hat{z}} & (95\text{-confidence level}) \\ c_{0.01} &= 2.58\sigma_{\hat{z}} & (99\text{-confidence level}) \\ c_{0.001} &= 3.29\sigma_{\hat{z}} & (99.9\text{-confidence level}) . \end{cases}$$

From this, the link with the standard deviation $\sigma_{\hat{z}}$ becomes clear. Here, the confidence interval of an estimator \hat{z} has been presented. However, similarly, the confidence interval of an observation y can be evaluated.

Instead of evaluating the confidence interval of a single variable, it is also possible to evaluate a confidence region of a random vector. Here, only the specific example will be considered that an estimator \hat{z} resides in an ellipsoidal region with its size governed by the confidence coefficient c_α . The reason is that

$$\|\hat{z} - z\|_{Q_{\hat{z}}}^2 = c_\alpha$$

is the equation of an hyper-ellipsoid and if

$$\hat{z} \sim N(z, Q_{\hat{z}}) ,$$

then

$$\|\hat{z} - z\|_{Q_{\hat{z}}}^2 \sim \chi^2(p, 0) , \quad (22.30)$$

where $\chi^2(p, 0)$ denotes the central chi-square distribution with p degrees of freedom, and p is the dimension of vector \hat{z} . This allows to evaluate the probability

$$P(\|\hat{z} - z\|_{Q_{\hat{z}}}^2 < c_\alpha) = 1 - \alpha \quad (22.31)$$

and again c_α can be determined for a given choice of α (Fig. 22.3b).

Example 22.1 Confidence Ellipse for Horizontal Positioning

With GNSS positioning, the main parameters of interest are usually the position coordinates. It can then be particularly useful to consider the ellipsoidal confidence region as presented here. See the example in Fig. 22.4, where only the horizontal positioning errors are considered.

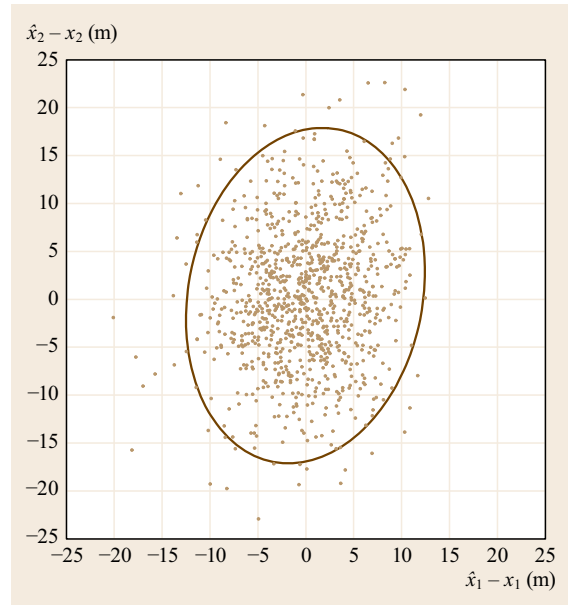


Fig. 22.4 95%-confidence ellipse for the horizontal positioning errors, where x_1 is the true east coordinate, and x_2 the true north coordinate. The brown dots represent the actual errors based on 1000 estimated positions (simulated)

22.3 Special Forms of Least Squares

In this section, several special forms of LS estimation will be discussed [22.2]. First, various partitioned models and their solutions are reviewed. In the special case that certain constraints on the unknown parameters can be applied, the problem of constrained LS estimation arises, which is the topic of Sect. 22.3.4. The problem of rank-deficiency, causing the solution to be nonunique, is briefly discussed in Sect. 22.3.5. Finally, the problem of nonlinear LS is addressed in Sect. 22.3.6. Another special form of LS arises for the mixed-integer model (Chap. 23), that is, if a subset of the parameters must fulfill an integer constraint. A well-known example is the GNSS model with integer carrier-phase ambiguities.

22.3.1 Recursive Estimation

A common form of a partitioned model is

$$E \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_k \end{bmatrix} \mathbf{x} \quad (22.32)$$

with

$$\mathbf{Q}_{yy} = \begin{bmatrix} \mathbf{Q}_1 & & & \mathbf{0} \\ & \mathbf{Q}_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{Q}_k \end{bmatrix}, \quad (22.33)$$

where the observation vectors y_i correspond to subsequent epochs $i = 1, \dots, k$.

The BLUE of \mathbf{x} using all k epochs of data follows as

$$\hat{\mathbf{x}}_{(k)} = \left(\sum_{i=1}^k \mathbf{A}_i^\top \mathbf{Q}_i^{-1} \mathbf{A}_i \right)^{-1} \left(\sum_{i=1}^k \mathbf{A}_i^\top \mathbf{Q}_i^{-1} y_i \right). \quad (22.34)$$

The model can be solved recursively, that is, a new solution is computed for every new epoch of data, as follows. At epoch 1, the model $E(y_1) = \mathbf{A}_1 \mathbf{x}$ with $\mathbf{Q}_{yy} = \mathbf{Q}_1$ is solved as

$$\begin{aligned} \hat{\mathbf{x}}_{(1)} &= (\mathbf{A}_1^\top \mathbf{Q}_1^{-1} \mathbf{A}_1)^{-1} \mathbf{A}_1^\top \mathbf{Q}_1^{-1} y_1, \\ \mathbf{Q}_{\hat{\mathbf{x}}_{(1)}} &= (\mathbf{A}_1^\top \mathbf{Q}_1^{-1} \mathbf{A}_1)^{-1}. \end{aligned} \quad (22.35)$$

In the following epochs, it is not necessary to apply a so-called *batch* LS solution by jointly processing all measurement data up till and including the current

epoch. Instead, the following model can be used at epoch $k = 2, 3, \dots$

$$E \begin{pmatrix} \hat{\mathbf{x}}_{(k-1)} \\ y_k \end{pmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{A}_k \end{bmatrix} \mathbf{x}; \begin{bmatrix} \mathbf{Q}_{\hat{\mathbf{x}}_{(k-1)}} & \\ & \mathbf{Q}_k \end{bmatrix} \quad (22.36)$$

with solution

$$\hat{\mathbf{x}}_{(k)} = \mathbf{Q}_{\hat{\mathbf{x}}_{(k)}}^{-1} \left(\mathbf{Q}_{\hat{\mathbf{x}}_{(k-1)}}^{-1} \hat{\mathbf{x}}_{(k-1)} + \mathbf{A}_k^\top \mathbf{Q}_k^{-1} y_k \right) \quad (22.37)$$

and

$$\mathbf{Q}_{\hat{\mathbf{x}}_{(k)}} = \left(\mathbf{Q}_{\hat{\mathbf{x}}_{(k-1)}}^{-1} + \mathbf{A}_k^\top \mathbf{Q}_k^{-1} \mathbf{A}_k \right)^{-1}$$

with

$$\mathbf{Q}_{\hat{\mathbf{x}}_{(k-1)}}^{-1} = \sum_{i=1}^{k-1} \mathbf{A}_i^\top \mathbf{Q}_i^{-1} \mathbf{A}_i.$$

The recursive BLUE solution of (22.37) can also be written as

$$\hat{\mathbf{x}}_{(k)} = \hat{\mathbf{x}}_{(k-1)} + \mathbf{K}_k (y_k - \mathbf{A}_k \hat{\mathbf{x}}_{(k-1)}), \quad k > 1 \quad (22.38)$$

with gain matrix

$$\mathbf{K}_k = \left(\mathbf{Q}_{\hat{\mathbf{x}}_{(k-1)}}^{-1} + \mathbf{A}_k^\top \mathbf{Q}_k^{-1} \mathbf{A}_k \right)^{-1} \mathbf{A}_k^\top \mathbf{Q}_k^{-1}. \quad (22.39)$$

Equation (22.38) is referred to as the *measurement update* (MU), since the right-hand side of this equation comprises of the estimator based on all previous epochs, $\hat{\mathbf{x}}_{(k-1)}$, plus the term $\mathbf{K}_k (y_k - \mathbf{A}_k \hat{\mathbf{x}}_{(k-1)})$ in which $\mathbf{A}_k \hat{\mathbf{x}}_{(k-1)}$ can be interpreted as the prediction of y_k . The difference $\mathbf{v}_k = y_k - \mathbf{A}_k \hat{\mathbf{x}}_{(k-1)}$ is consequently called the *predicted residual*. Matrix \mathbf{K}_k is called the gain matrix, since the gain experienced by the new data y_k is determined by $\mathbf{K}_k \mathbf{v}_k$.

An alternative form of the gain matrix is

$$\mathbf{K}_k = \mathbf{Q}_{\hat{\mathbf{x}}_{(k-1)}} \mathbf{A}_k^\top (\mathbf{Q}_k + \mathbf{A}_k \mathbf{Q}_{\hat{\mathbf{x}}_{(k-1)}} \mathbf{A}_k^\top)^{-1}. \quad (22.40)$$

Based on this form, the variance–covariance matrix can be calculated with

$$\mathbf{Q}_{\hat{\mathbf{x}}_{(k)}} = (\mathbf{I} - \mathbf{K}_k \mathbf{A}_k) \mathbf{Q}_{\hat{\mathbf{x}}_{(k-1)}}. \quad (22.41)$$

The results obtained with (22.39) and (22.40) will be identical. Preference for one of the two expressions may be based on the dimensions of the inverse matrices. If the number of observations m_k is small compared to

the state vector dimension n , one may prefer (22.40) over (22.39).

Recursive estimation can be applied in real-time estimation problems, providing new estimates at every measurement epoch using the information of subsequent epochs. For post-processing applications, the data collected over a longer time period can alternatively be jointly processed with batch LS to find the estimates that provide the best fit to the entire set of measurements.

Example 22.2 (Single Point Positioning)

An example is shown in Fig. 22.5. The solutions based on single epoch processing, that is, not using data from previous epochs, are shown as dots for a single point positioning model (Sect. 21.3). The recursive LS solution is shown with the solid lines. It can be seen that the solution, especially for the horizontal position, converges quite fast, whereas the precision of the single epoch solutions is much poorer (larger spread in outcomes). Note that the batch solution based on all epochs is identical to the recursive solution at the last epoch.

In Sect. 22.5, another step, namely a time update, will be included to account for a kinematic or dynamic model, that is, with \mathbf{x} not being constant in time.

Example 22.3 (Recursive Phase-Smoothed Pseudorange)

The phase-smoothed pseudorange algorithm is a single-channel recursive algorithm that uses the high-precision carrier-phase data ϕ_k to *smooth* the relatively high noise of the pseudorange data p_k . The smoothed pseudorange

$\hat{p}_{k|k}$ is obtained from ϕ_k and p_k as [22.7, 8]

$$\begin{aligned}\hat{p}_{k|k-1} &= \hat{p}_{k-1|k-1} + [\phi_k - \phi_{k-1}], \\ \hat{p}_{k|k} &= \hat{p}_{k|k-1} + \frac{1}{k} [p_k - \hat{p}_{k|k-1}].\end{aligned}\quad (22.42)$$

The algorithm is initialized with $\hat{p}_{1|1} = p_1$. The two equations can be combined to give

$$\hat{p}_{k|k} = \frac{1}{k} p_k + \frac{k-1}{k} [\hat{p}_{k-1|k-1} + (\phi_k - \phi_{k-1})]. \quad (22.43)$$

This equation shows that the *smoothed* pseudorange is a linear combination of the pseudorange, with weight $1/k$, and the *predicted* pseudorange, with weight $k-1/k$. Figure 22.6 shows a time series of the noise of p_k and $\hat{p}_{k|k}$, respectively.

The following example shows how the smoothed pseudorange algorithm is linked to recursive LS estimation.

Example 22.4 (Recursive Phase-Adjusted Pseudorange)

Let the multi-epoch, kinematic GNSS model of observation equations be given as

$$E \begin{pmatrix} p_1 \\ \phi_1 \\ p_2 \\ \phi_2 \\ \vdots \\ p_k \\ \phi_k \end{pmatrix} = \begin{bmatrix} \mathbf{A}_1 & & & \mathbf{0} \\ \mathbf{A}_1 & & & \mathbf{I} \\ & \mathbf{A}_2 & & \mathbf{0} \\ & \mathbf{A}_2 & & \mathbf{I} \\ & & \ddots & \vdots \\ & & & \mathbf{A}_k \\ & & & \mathbf{A}_k & \mathbf{I} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ a \end{pmatrix} \quad (22.44)$$

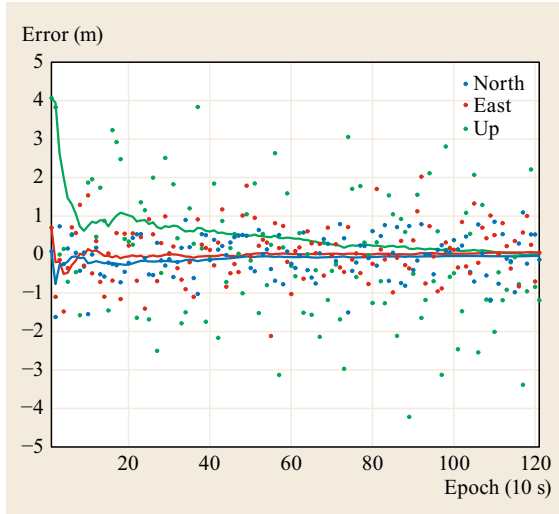


Fig. 22.5 Single epoch (dots) and recursive least-squares (lines) solutions with single point positioning

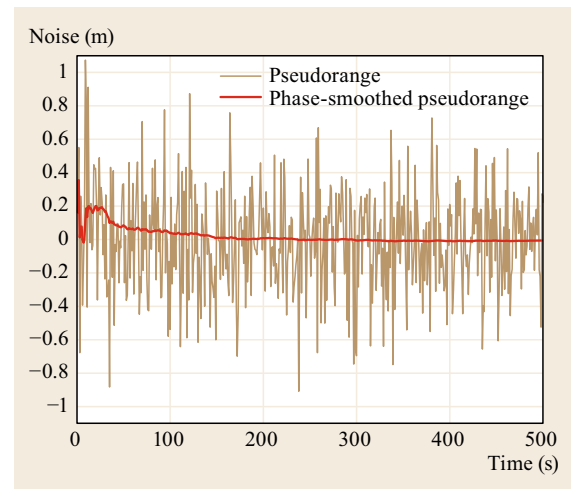


Fig. 22.6 Noise on pseudorange observation p_k (brown) and phase-smoothed pseudorange observation $\hat{p}_{k|k}$ (red)

with p_i and ϕ_i the vectors of (observed minus computed) pseudorange and phase observables at epoch i , respectively; x_i the vector of (increment) position coordinates, clock errors, and a the unknown ambiguities (in meters), which are assumed to be constant in time in the absence of cycle slips. The atmospheric delays are assumed corrected for and/or cancelled by means of using atmosphere-free observables.

The recursive LS solution of model (22.44) is given as [22.9]

$$\begin{aligned}\hat{x}_k &= Q_{\hat{x}_k \hat{x}_k} A_k^\top \left(Q_{p_k p_k}^{-1} p_k + Q_{\bar{p}_k \bar{p}_k}^{-1} \bar{p}_k \right) \\ a_{(k)} &= \hat{a}_{(k-1)} + Q_{\hat{a} \hat{a}_{(k-1)}} Q_{\bar{p}_k \bar{p}_k}^{-1} (\bar{p}_k - A_k \hat{x}_k)\end{aligned}\quad (22.45)$$

with

$$\begin{aligned}Q_{\hat{x}_k \hat{x}_k} &= \left(A_k^\top \left(Q_{p_k p_k}^{-1} + Q_{\bar{p}_k \bar{p}_k}^{-1} \right) A_k \right)^{-1}, \\ \bar{p}_k &= \phi_k - \hat{a}_{(k-1)}, \quad Q_{\bar{p}_k \bar{p}_k} = Q_{\phi_k \phi_k} + Q_{\hat{a} \hat{a}_{(k-1)}}.\end{aligned}$$

The phase-adjusted pseudorange estimator is given by $A_k \hat{x}_k$.

If we now make the simplifying assumption that $Q_{\phi_k \phi_k} = 0$, then $Q_{\bar{p}_k \bar{p}_k} = Q_{\hat{a} \hat{a}_{(k-1)}}$ and $\hat{a}_{(k)} = \phi_k - A_k \hat{x}_k$, from which it follows that (22.45) simplifies to

$$\hat{x}_k = K_k p_k + L_k [A_{k-1} \hat{x}_{k-1} + (\phi_k - \phi_{k-1})] \quad (22.46)$$

with $K_k = Q_{\hat{x}_k \hat{x}_k} A_k^\top Q_{p_k p_k}^{-1}$, $L_k = Q_{\hat{x}_k \hat{x}_k} A_k^\top Q_{\hat{a} \hat{a}_{k-1}}^{-1}$. Compare (22.46) with (22.43). The result (22.46) becomes even identical to (22.43) in case of single-channel processing. Then, the relative-receiver satellite geometry is absent from the model and thus $A_k = I$, from which it follows that $K_k = \frac{1}{k} I$ and $L_k = \frac{k-1}{k} I$. The phase-smoothed pseudorange estimator thus becomes a LS estimator when the phase noise is neglected and the processing is restricted to single-channel processing.

The phase-adjusted pseudorange estimator, being a BLUE, has a better precision than that of the phase-smoothed estimator. The difference in precision between the two estimators, that is, their variance difference, is given in [22.9] as

$$\frac{(k-1)\sigma_\phi^2/\sigma_p^2}{k \left(1 + \sigma_\phi^2/\sigma_p^2 \right)} \sigma_\phi^2. \quad (22.47)$$

This difference is zero at initialization ($k=1$) and nonzero but very small otherwise ($\sigma_\phi^2 \ll \sigma_p^2$). Hence, when restricted to single-channel processing, the phase-smoothed pseudorange is very close to the optimal phase-adjusted pseudorange estimator.

22.3.2 Estimation with Partitioned Parameter Vector

The model with the unknown parameter vector x partitioned can be written as

$$E(y) = [A_1 \quad A_2] \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \quad (22.48)$$

This partitioning can be relevant in case one is interested in only a subset of the unknown parameters. The BLUE of, for example, the subset x_1 equals

$$\hat{x}_1 = \underbrace{(\bar{A}_1^\top Q_{yy}^{-1} \bar{A}_1)^{-1} \bar{A}_1^\top Q_{yy}^{-1} y}_{N_{\text{red}}} \quad (22.49)$$

with N_{red} the *reduced* normal matrix, reduced for x_2 , and

$$Q_{\hat{x}_1 \hat{x}_1} = N_{\text{red}}^{-1} \quad (22.50)$$

with $\bar{A}_1 = P_{A_2}^\perp A_1$.

Once \hat{x}_1 is known, it is possible to find \hat{x}_2 using

$$\hat{x}_2 = (A_2^\top Q_{yy}^{-1} A_2)^{-1} A_2^\top Q_{yy}^{-1} (y - A_1 \hat{x}_1) \quad (22.51)$$

and

$$\begin{aligned}Q_{\hat{x}_2 \hat{x}_2} &= (A_2^\top Q_{yy}^{-1} A_2)^{-1} \\ &\times \left(I + A_2^\top Q_{yy}^{-1} A_1 Q_{\hat{x}_1 \hat{x}_1} A_1^\top Q_{yy}^{-1} A_2 (A_2^\top Q_{yy}^{-1} A_2)^{-1} \right).\end{aligned}\quad (22.52)$$

The roles of x_1 and x_2 can of course be interchanged in (22.49)–(22.52).

Example 22.5 (Time-Varying and Time-Constant Parameters)

An example when the partitioned model (22.48) applies is for GNSS models comprising time-varying and time-constant parameters. For instance, for a static positioning application, the observations of k epochs are collected; the position and possibly ambiguity parameters remain constant in time and are parameterized in vector x_1 , whereas the clock and atmospheric parameters are time-varying and parameterized in vector x_2 . The structure of matrices A_1 and A_2 is then

$$A_1 = \begin{bmatrix} A_{11} \\ \vdots \\ A_{1k} \end{bmatrix}, \quad A_2 = \begin{bmatrix} A_{21} & & \\ & \ddots & \\ & & A_{2k} \end{bmatrix}. \quad (22.53)$$

22.3.3 Block Estimation

A partitioned model often arising when observations from many stations or two or more (geodetic) networks, campaigns, or sessions are combined as

$$E \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_{1g} \\ \mathbf{0} & \mathbf{A}_{22} & & \vdots & \mathbf{A}_{2g} \\ \vdots & & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_{kk} & \mathbf{A}_{kg} \end{bmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ x_g \end{pmatrix}, \quad (22.54)$$

where k is the number of, for example, stations/networks. The vector x_g contains the global parameters (related to all y_i), and x_i ($i = 1, \dots, k$) the local parameter vector related to observations y_i only. The observations y_i are assumed to be mutually independent

$$\mathbf{Q}_{yy} = \begin{bmatrix} \mathbf{Q}_1 & & & \mathbf{0} \\ & \mathbf{Q}_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{Q}_k \end{bmatrix}. \quad (22.55)$$

The solution can be obtained with the *Helmert block method* [22.2, 10]. First, the global parameters are estimated as

$$\hat{x}_g = \left(\underbrace{\sum_{i=1}^k \bar{\mathbf{A}}_{ig}^T \mathbf{Q}_i^{-1} \bar{\mathbf{A}}_{ig}}_{\mathbf{N}_{\text{red}}} \right)^{-1} \left(\sum_{i=1}^k \bar{\mathbf{A}}_{ig}^T \mathbf{Q}_i^{-1} y_i \right). \quad (22.56)$$

\mathbf{N}_{red} is the reduced normal matrix, since x_i are reduced, with

$$\bar{\mathbf{A}}_{ig} = \mathbf{P}_{\mathbf{A}_{ii}}^\perp \mathbf{A}_{ig}, \quad (22.57)$$

and the orthogonal projector is given as $\mathbf{P}_{\mathbf{A}_{ii}}^\perp = \mathbf{I} - \mathbf{P}_{\mathbf{A}_{ii}}$ and $\mathbf{P}_{\mathbf{A}_{ii}} = \mathbf{A}_{ii}(\mathbf{A}_{ii}^T \mathbf{Q}_i^{-1} \mathbf{A}_{ii})^{-1} \mathbf{A}_{ii}^T \mathbf{Q}_i^{-1}$. Using this solution, the local parameters can be estimated

$$\hat{x}_i = (\mathbf{A}_{ii}^T \mathbf{Q}_i^{-1} \mathbf{A}_{ii})^{-1} \mathbf{A}_{ii}^T \mathbf{Q}_i^{-1} (y_i - \mathbf{A}_{ig} \hat{x}_g). \quad (22.58)$$

Example 22.6 (Large GNSS Network)

Rigorous LS estimation of precise orbits and clocks

based on observations from around 20 000 GNSS receivers is hampered by the processing capacity of the IGS analysis centers [22.11]. The Helmert block method can be applied to deal with this problem. In this case, the vectors y_i refer to the observation vectors from the individual GNSS stations. The global parameters x_g are the satellite orbit and clock parameters, and Earth rotation parameters. The local parameters x_i are the station coordinates, atmospheric delays, and receiver clock errors per station i .

Similar applications of the Helmert block method are described in [22.12, 13]. An application for the readjustment of a large national geodetic network is described in [22.14].

22.3.4 Constrained Least-Squares

So far we discussed the unconstrained linear model $E(y) = \mathbf{A}x$. Another formulation is used if certain constraints on the parameters are to be considered

$$E(y) = \mathbf{A}x \quad \text{with} \quad \mathbf{C}^T x = c. \quad (22.59)$$

This is the constrained linear model. Compared to the unconstrained model, the system of constraints $\mathbf{C}^T x = c$ are added with $n \times d$ matrix \mathbf{C} of rank d , and d -vector c . The redundancy of this model is increased with d compared to the unconstrained model.

In order to solve the model (22.59), first the unconstrained solution, denoted as \hat{x} and $\mathbf{Q}_{\hat{x}\hat{x}}$, is computed. In the second step, the conditioned linear model $\mathbf{C}^T E(\hat{x}) = c$ is solved as

$$\hat{x}_c = \hat{x} - \mathbf{Q}_{\hat{x}\hat{x}} \mathbf{C} (\mathbf{C}^T \mathbf{Q}_{\hat{x}\hat{x}} \mathbf{C})^{-1} (\mathbf{C}^T \hat{x} - c). \quad (22.60)$$

The corresponding variance-covariance matrix is obtained by applying the variance propagation law

$$\mathbf{Q}_{\hat{x}_c \hat{x}_c} = \mathbf{Q}_{\hat{x}\hat{x}} - \mathbf{Q}_{\hat{x}\hat{x}} \mathbf{C} (\mathbf{C}^T \mathbf{Q}_{\hat{x}\hat{x}} \mathbf{C})^{-1} \mathbf{C}^T \mathbf{Q}_{\hat{x}\hat{x}}. \quad (22.61)$$

22.3.5 Rank-Defect Least Squares

No unique LS solution exists if the design matrix \mathbf{A} fails to be of full rank.

Let $\text{rank}(\mathbf{A}) = r < n$. Then, an $n \times (n-r)$ basis matrix \mathbf{V} exists with the property $\mathbf{A}\mathbf{V} = \mathbf{0}$. Hence, no unique LS-solution exists then for

$$\min_x \|y - \mathbf{A}x\|_{\mathbf{Q}_{yy}}^2,$$

since if, say \hat{x}_s , is an LS-solution, then so is $\hat{x}_s + \mathbf{V}\beta$ for any $\beta \in \mathbb{R}^{n-r}$.

To determine the general solution of a rank-defect LS problem, we first determine a particular solution. Let \mathbf{S} be a basis matrix of a subspace complementary to the range space of \mathbf{V} , that is, $\mathcal{R}(\mathbf{S}) \oplus \mathcal{R}(\mathbf{V}) = \mathbb{R}^n$. Then, \mathbf{S} is of order $n \times r$ and matrix $[\mathbf{S}, \mathbf{V}]$ is invertible. With the reparameterization

$$\mathbf{x} = \mathbf{S}\boldsymbol{\alpha} + \mathbf{V}\boldsymbol{\beta} \quad (22.62)$$

the originally rank deficient system is turned into a full rank system

$$E(\mathbf{y}) = \mathbf{A}\mathbf{x} \Rightarrow E(\mathbf{y}) = (\mathbf{A}\mathbf{S})\boldsymbol{\alpha} \quad (22.63)$$

having the LS-solution

$$\hat{\boldsymbol{\alpha}} = [(\mathbf{A}\mathbf{S})^\top \mathbf{Q}_{yy}^{-1} (\mathbf{A}\mathbf{S})]^{-1} (\mathbf{A}\mathbf{S})^\top \mathbf{Q}_{yy}^{-1} \mathbf{y}.$$

Hence, $\hat{\mathbf{x}}_s = \mathbf{S}\hat{\boldsymbol{\alpha}}$ is a particular solution of the rank-defect LS-problem, while its general LS-solution is given by

$$\hat{\mathbf{x}} = \hat{\mathbf{x}}_s + \mathbf{V}\boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^{n-r}. \quad (22.64)$$

Instead of using matrix \mathbf{S} to compute $\hat{\mathbf{x}}_s$, one may also use a computation that makes use of a basis matrix of the orthogonal complement of $\mathcal{R}(\mathbf{S})$. If the $n \times (n-r)$ matrix \mathbf{S}^\perp is such a basis matrix, that is, $\mathcal{R}(\mathbf{S}^\perp) = \mathcal{R}(\mathbf{S})^\perp$, then $\hat{\mathbf{x}}_s$ is the LS-solution of $E(\mathbf{y}) = \mathbf{A}\mathbf{x}$ with constraints $\mathbf{S}^{\perp\top} \mathbf{x} = \mathbf{0}$ [22.15].

In decomposition (22.62), $\tilde{\mathbf{x}}_s = \mathbf{S}\boldsymbol{\alpha}$ contains the estimable parameter part, whereas $\mathbf{x}_v = \mathbf{V}\boldsymbol{\beta}$ contains the undetermined or inestimable part (Fig. 22.7)

$$\mathbf{x} = \tilde{\mathbf{x}}_s + \mathbf{x}_v, \tilde{\mathbf{x}}_s \in \mathcal{R}(\mathbf{S}), \mathbf{x}_v \in \mathcal{R}(\mathbf{V}). \quad (22.65)$$

Thus, $\hat{\mathbf{x}}_s$ is *not* an unbiased estimator of \mathbf{x} , $E(\hat{\mathbf{x}}_s) \neq \mathbf{x}$, but of $\tilde{\mathbf{x}}_s$ instead, $E(\hat{\mathbf{x}}_s) = \tilde{\mathbf{x}}_s$. To understand what $\hat{\mathbf{x}}_s$ estimates, we need the relation between $\tilde{\mathbf{x}}_s$ and \mathbf{x} . It is given by

$$\tilde{\mathbf{x}}_s = \mathbf{S}\mathbf{x}, \text{ with } \mathbf{S} = \mathbf{I}_n - \mathbf{V}[\mathbf{S}^{\perp\top} \mathbf{V}]^{-1} \mathbf{S}^{\perp\top}. \quad (22.66)$$

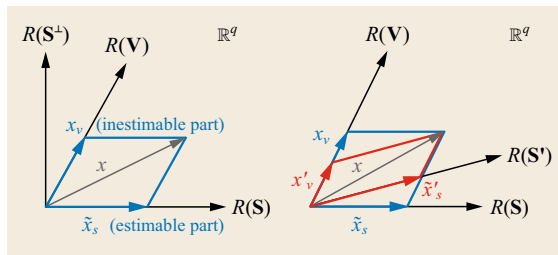


Fig. 22.7 Two examples of decomposing \mathbf{x} in an estimable and inestimable part: $\mathbf{x} = \tilde{\mathbf{x}}_s + \mathbf{x}_v = \tilde{\mathbf{x}}'_s + \mathbf{x}'_v$. The S -transformation, transforms between solutions: $\tilde{\mathbf{x}}_s = \mathbf{S}\tilde{\mathbf{x}}'_s$ (after [22.16])

This relation shows which linear functions of \mathbf{x} are estimated by $\hat{\mathbf{x}}_s$. Matrix \mathbf{S} , referred to as S (ingularity)-transformation, is a projector (i.e., an idempotent matrix) that projects onto $\mathcal{R}(\mathbf{S})$ and along $\mathcal{R}(\mathbf{V})$ [22.15, 17]. This projection is depicted in Fig. 22.7. With the use of the S -transformation, one can obtain any particular solution, say $\hat{\mathbf{x}}_s$ and its variance-covariance matrix $\mathbf{Q}_{\hat{\mathbf{x}}_s \hat{\mathbf{x}}_s}$, from any member of the general solution $\hat{\mathbf{x}}$ and its variance-covariance matrix, as

$$\hat{\mathbf{x}}_s = \mathbf{S}\hat{\mathbf{x}}, \quad \mathbf{Q}_{\hat{\mathbf{x}}_s \hat{\mathbf{x}}_s} = \mathbf{S}\mathbf{Q}_{\hat{\mathbf{x}} \hat{\mathbf{x}}} \mathbf{S}^\top. \quad (22.67)$$

Hence, one can also transform between different particular solutions by means of an S -transformation.

22.3.6 Non-Linear Least-Squares

So far only linear LS has been considered. However, the LS principle can also be applied to nonlinear systems of equations. Consider therefore the following inconsistent system of equations

$$y_i \approx a_i(\mathbf{x}), \quad i = 1, \dots, m, \quad (22.68)$$

where $a_i : \mathbb{R}^n \mapsto \mathbb{R}$ are known but nonlinear functions of the unknown parameter vector \mathbf{x} . In compact form this is written as

$$\mathbf{y} \approx \mathbf{A}(\mathbf{x}) = \begin{bmatrix} a_1(\mathbf{x}) \\ \vdots \\ a_m(\mathbf{x}) \end{bmatrix}. \quad (22.69)$$

The corresponding nonlinear WLS solution reads then

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{y} - \mathbf{A}(\mathbf{x})\|_{\mathbf{W}^{-1}}^2. \quad (22.70)$$

This problem can be solved using an iterative procedure, starting with an initial approximation \mathbf{x}_0 of the unknown parameters. Successive approximations are made based on the linearized version of $\mathbf{A}(\mathbf{x})$.

First, a Taylor series expansion of $a_i(\mathbf{x})$ is used

$$a_i(\mathbf{x}) = a_i(\mathbf{x}_0) + [\partial_x a_i(\mathbf{x}_0)]^\top (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\top \mathbf{H}(\boldsymbol{\theta})(\mathbf{x} - \mathbf{x}_0), \quad (22.71)$$

where $a_i : \mathbb{R}^n \mapsto \mathbb{R}$ must have continuous second-order partial derivatives. The gradient vector is given as

$$\partial_x a_i(\mathbf{x}_0) = \begin{bmatrix} \frac{\partial}{\partial x_1} a_i(\mathbf{x}_0) \\ \vdots \\ \frac{\partial}{\partial x_n} a_i(\mathbf{x}_0) \end{bmatrix}$$

and the Hessian matrix

$$\mathbf{H}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} a_i(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial x_1 \partial x_n} a_i(\boldsymbol{\theta}) \\ \vdots & & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} a_i(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial x_n \partial x_n} a_i(\boldsymbol{\theta}) \end{bmatrix}$$

with $\boldsymbol{\theta}$ in between \mathbf{x} and \mathbf{x}_0 . The last term of (22.71) is the second-order remainder and will be ignored, such that the linear approximation

$$a_i(\mathbf{x}) \approx a_i(\mathbf{x}_0) + [\partial_{\mathbf{x}} a_i(\mathbf{x}_0)]^\top (\mathbf{x} - \mathbf{x}_0)$$

is obtained. The vector function $\mathbf{A}(\mathbf{x})$ can now be approximated as

$$\mathbf{A}(\mathbf{x}) \approx \mathbf{A}(\mathbf{x}_0) + \mathbf{J}_0(\mathbf{x} - \mathbf{x}_0) \quad (22.72)$$

with the Jacobian matrix

$$\mathbf{J}_0 = \begin{bmatrix} [\partial_{\mathbf{x}} a_1(\mathbf{x}_0)]^\top \\ \vdots \\ [\partial_{\mathbf{x}} a_m(\mathbf{x}_0)]^\top \end{bmatrix}.$$

The nonlinear system (22.69) can be approximated using (22.72) as

$$\Delta \mathbf{y}_0 \approx \mathbf{J}_0 \Delta \mathbf{x}_0 \quad (22.73)$$

with $\Delta \mathbf{y}_0 = \mathbf{y} - \mathbf{A}(\mathbf{x}_0)$ and $\Delta \mathbf{x}_0 = \mathbf{x} - \mathbf{x}_0$.

The linearized system of equation in (22.73) can now be solved using the linear WLS principle to obtain the estimator

$$\Delta \hat{\mathbf{x}}_0 = (\mathbf{J}_0^\top \mathbf{W} \mathbf{J}_0)^{-1} \mathbf{J}_0^\top \mathbf{W} \Delta \mathbf{y}_0. \quad (22.74)$$

The WLS solution $\hat{\mathbf{x}}_0 = \mathbf{x}_0 + \Delta \hat{\mathbf{x}}_0$ should ideally be closer to \mathbf{x} than \mathbf{x}_0 but is generally not used as the final solution. In the so-called Gauss–Newton iteration scheme, this solution is namely used as a new approximation in order to get a better approximation and continues until the difference between successive approximations becomes small enough. The procedure is thus to take as successive approximations

$$\mathbf{x}_i := \mathbf{x}_{i-1} + \Delta \hat{\mathbf{x}}_{i-1} \quad (22.75)$$

with

$$\Delta \hat{\mathbf{x}}_i = (\mathbf{J}_i^\top \mathbf{W} \mathbf{J}_i)^{-1} \mathbf{J}_i^\top \mathbf{W} \Delta \mathbf{y}_i, \quad (22.76)$$

where \mathbf{J}_i is the matrix with partial derivatives with respect to \mathbf{x}_i and $\Delta \mathbf{y}_i = \mathbf{y} - \mathbf{A}(\mathbf{x}_i)$.

The iteration is terminated when

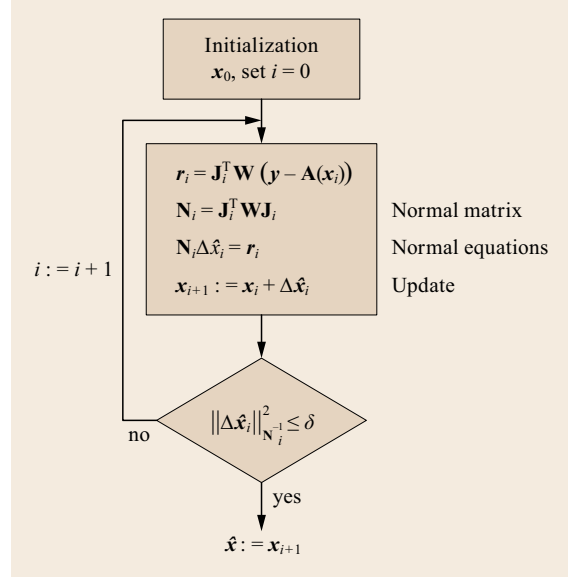


Fig. 22.8 Gauss–Newton iteration for nonlinear least-squares

$$\|\Delta \hat{\mathbf{x}}_i\|_{\mathbf{N}_i^{-1}}^2 \leq \delta \quad \text{with} \quad \mathbf{N}_i = \mathbf{J}_i^\top \mathbf{W} \mathbf{J}_i$$

and δ a small user-defined threshold. Once the stop criterion is met, the WLS solution is given by

$$\hat{\mathbf{x}} = \mathbf{x}_i + \Delta \hat{\mathbf{x}}_i. \quad (22.77)$$

The complete Gauss–Newton iteration scheme is depicted in Fig. 22.8. For more properties of the Gauss–Newton method and on alternative nonlinear-LS solvers, see [22.18, 19].

The initial approximation \mathbf{x}_0 has to be obtained depending on the application at hand; it requires insight in the specific problem.

The Gauss–Newton method will converge to the LS solution in the absence of severe nonlinearity and large measurement errors. However, it should be noted that the nonlinear WLS estimator is not unbiased and is not normally distributed even if the measurements and random errors are. Diagnostic measures to test for the significance of these nonlinearity effects on the mean and variance–covariance matrix are given in [22.20]. If the second- and higher order terms in the Taylor series expansion are negligibly small, the distribution of the nonlinear WLS estimator will be very close to the normal distribution with negligible bias. The variance–covariance matrix $\mathbf{Q}_{\hat{\mathbf{x}}}$ can then be approximated by

$$\mathbf{Q}_{\hat{\mathbf{x}}} \approx (\mathbf{J}_i^\top \mathbf{Q}_{yy}^{-1} \mathbf{J}_i)^{-1}, \quad (22.78)$$

where \mathbf{A} in (22.22) is replaced by the Jacobian \mathbf{J}_i (from the last iteration).

22.4 Prediction and Filtering

In most GNSS applications, the parameters of interest are time-varying. This is obvious for a moving platform, for which we are interested in its position and sometimes also its attitude (Chaps. 27 and 28). Also, with GNSS deformation measurements of the Earth's surface or man-made structures we are dealing with time-variable parameters. Another example is when we are interested in the troposphere and ionosphere delays, as well as instrumental parameters such as clock drifts and biases.

In Sect. 22.3.1, we have already seen how recursive estimation can be applied in case of constant model parameters with sequentially collected data. Recursive or sequential estimation is very suitable for real-time estimation problems, but also for offline processing in order to reduce the number of parameters to be simultaneously estimated. In the following, three types of estimation problems for sequentially collected data will be discussed. The estimation problem arising for real-time applications is referred to as *filtering*: The measurements up till and including the current epoch are used to estimate the current state. This is the topic of Sect. 22.5. There, also *smoothing* will be discussed. This is the problem where the time of interest falls within the time span of collected data. In the current chapter, the focus will be on *prediction*, this is when we are interested in a future state. Figure 22.9 illustrates the three problems.

22.4.1 Prediction Problem

Prediction may refer to predicting the model parameters in time; however, it may also refer to predicting, or interpolating the parameters in space. Moreover, not necessarily, the model parameters are to be predicted; also the observations themselves or the signal and noise affecting the measurements may have to be predicted. In all cases, the idea is that an observable random vector y is used to guess the outcome of another random vec-

tor, which is not observable. Note the difference with the estimation problem, where the random y was used to find the best matching unknown but *deterministic* parameters according to a given relationship.

All prediction problems mentioned above can be written in a similar form as (22.3) [22.21]

$$\begin{pmatrix} y \\ p \end{pmatrix} = \begin{bmatrix} A \\ A_p \end{bmatrix} x + \begin{pmatrix} e \\ e_p \end{pmatrix} \quad (22.79)$$

but now with p being an unobserved random vector of which the outcome needs to be predicted. As before, the unknown parameter vector x is deterministic, A_p is the given $m_p \times n$ coefficient matrix of p , and e_p is a zero-mean random vector. The variance-covariance matrix of y and p

$$\begin{bmatrix} Q_{yy} & Q_{yp} \\ Q_{py} & Q_{pp} \end{bmatrix}$$

is assumed given.

Some typical examples to which the above formulation applies are:

Example 22.7 Recovery of Undifferenced Residuals

Let e in $y = Ax + e$ be given as $e = D^T u$, with matrix D known and where u is a zero-mean random vector with known variance-covariance matrix Q_{uu} . If y is the vector of differenced GNSS observables and D^T the differencing operator, then u is the vector of undifferenced residuals. If the undifferenced GNSS residual vector u is to be predicted, then (22.79) takes the form

$$\begin{pmatrix} y \\ u \end{pmatrix} = \begin{bmatrix} A \\ 0 \end{bmatrix} x + \begin{pmatrix} e \\ u \end{pmatrix} \quad (22.80)$$

with variance-covariance matrix

$$\begin{bmatrix} D^T Q_{uu} D & D^T Q_{uu} \\ Q_{uu} D & Q_{uu} \end{bmatrix}.$$

A typical GNSS example where the method of recovering undifferenced residuals from double-differenced data is applied is in the creation of CORS multipath maps [22.22, 23].

Example 22.8 Separation of Trend, Signal, and Noise

For the trend-signal-noise model, we have

$$y = Ax + s + n, \quad (22.81)$$

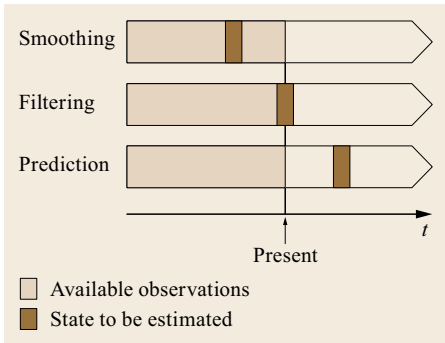


Fig. 22.9 The concept of smoothing, filtering, and prediction

where $\mathbf{A}\mathbf{x}$ is the deterministic *trend* with unknown \mathbf{x} , s is a zero-mean random *signal*, and \mathbf{n} is the zero-mean random *noise*. In this case, the goal is to predict s and \mathbf{n} , and model (22.79) becomes

$$\begin{pmatrix} y \\ s \\ \mathbf{n} \end{pmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \mathbf{x} + \begin{pmatrix} s + \mathbf{n} \\ s \\ \mathbf{n} \end{pmatrix}. \quad (22.82)$$

Alternatively, the trend-signal-noise model can be applied to predict a function of the same type, say $y_p = \mathbf{A}_p \mathbf{x} + s_p$, at another *time* and/or *location*, which means interpolation or extrapolation. For instance, in the case of ionospheric interpolation from network-derived ionospheric delays, \mathbf{x} contains the ionospheric trend parameters, the signals s and s_p the spatiotemporal variability of the ionosphere, and \mathbf{n} the GNSS observational measurement noise. In that case the prediction model becomes

$$\begin{pmatrix} y \\ y_p \end{pmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{A}_p \end{bmatrix} \mathbf{x} + \begin{pmatrix} s + \mathbf{n} \\ s_p \end{pmatrix}, \quad (22.83)$$

which is again of a form similar as (22.79).

In this section, the principle and properties of minimum mean squared error (MMSE) prediction will be presented in order to solve the prediction problems described here. The solution will be presented based on the model of (22.79).

22.4.2 Minimum Mean Squared Error Prediction

Best Predictor (BP)

Let the predictor of \mathbf{p} be given by a function $G(\mathbf{y})$. The prediction error in model (22.79) will then be equal to

$$\mathbf{p} - G(\mathbf{y}). \quad (22.84)$$

A predictor is considered best if it has the smallest mean square prediction error. The BP $\hat{G}(\mathbf{y})$ from a certain class Ω of predictors should thus fulfill according to the MMSE criterion

$$E(\|\mathbf{p} - \hat{G}(\mathbf{y})\|_{\mathbf{W}^{-1}}^2) = \min_{\hat{G} \in \Omega} E(\|\mathbf{p} - G(\mathbf{y})\|_{\mathbf{W}^{-1}}^2). \quad (22.85)$$

If no restrictions are placed on the class of predictors, the best predictor of \mathbf{p} is equal to the conditional mean

$$\hat{\mathbf{p}}_{\text{BP}} = E(\mathbf{p}|\mathbf{y}) = \int \alpha f_{p|\mathbf{y}}(\alpha|\mathbf{y}) d\alpha \quad (22.86)$$

with $f_{p|\mathbf{y}}(\cdot)$ the conditional PDF of \mathbf{p} .

Best Linear Predictor (BLP)

If it is required that the predictor G is of a linear form

$$G(\mathbf{y}) = \mathbf{L}\mathbf{y} + \mathbf{l}, \quad (22.87)$$

then the best linear predictor (BLP) takes the form

$$\hat{\mathbf{p}}_{\text{BLP}} = \bar{\mathbf{p}} + \mathbf{Q}_{py} \mathbf{Q}_{yy}^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \quad (22.88)$$

with $\bar{\mathbf{p}} = E(\mathbf{p})$ and $\bar{\mathbf{y}} = E(\mathbf{y})$. Note that the BLP does not need complete information about the PDF, this in contrast with the BP. For the BLP, only the first two moments, the mean and the variance, need to be known. The two predictors, BP and BLP, become identical in the Gaussian case, that is, when \mathbf{y} and \mathbf{p} are normally distributed.

Best Linear Unbiased Prediction (BLUP)

Although the BLP only needs knowledge about the first two moments, the need to know the mean is still impractical for many applications. As can be seen from (22.79), since \mathbf{x} is unknown, also the means $\bar{\mathbf{p}}$ and $\bar{\mathbf{y}}$ in (22.88) are unknown. What is known, however, is that their means are linearly related as

$$E \begin{pmatrix} y \\ \mathbf{p} \end{pmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{A}_p \end{bmatrix} \mathbf{x}. \quad (22.89)$$

If we include this relation into the minimization of the mean squared prediction error over the class of linear unbiased predictors, one obtains the best linear unbiased predictor as

$$\hat{\mathbf{p}}_{\text{BLUP}} = \mathbf{A}_p \hat{\mathbf{x}} + \mathbf{Q}_{py} \mathbf{Q}_{yy}^{-1} (\mathbf{y} - \mathbf{A} \hat{\mathbf{x}}), \quad (22.90)$$

in which $\hat{\mathbf{x}}$ is the BLUE of \mathbf{x} from (22.21). Hence, the BLUP (22.90) follows from the BP (22.88) by replacing the means $\bar{\mathbf{p}}$ and $\bar{\mathbf{y}}$ by their BLUEs.

Example 22.9 Recovery of Undifferenced Residuals (Continued)

If we apply the above prediction formula (22.90) to determine the BLUP of the undifferenced GNSS residual \mathbf{u} of (22.80), we obtain

$$\hat{\mathbf{u}} = \mathbf{Q}_{uu} \mathbf{D} (\mathbf{D}^\top \mathbf{Q}_{uu} \mathbf{D})^{-1} \hat{\mathbf{e}} \quad (22.91)$$

with $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}}$. This result shows how the undifferenced GNSS residual $\hat{\mathbf{u}}$ can be obtained from the differenced GNSS residual $\hat{\mathbf{e}}$ [22.22, 23].

Example 22.10 Example of Trend-Signal-Noise

Consider a random vector function $z(t)$ (Fig. 22.10). We assume that $z(t)$ can be written as the sum of a deterministic, but unknown trend $A_t x$, and a zero-mean random signal $s(t)$

$$z(t) = A_t x + s(t), \quad (22.92)$$

where A_t is a known $k \times n$ matrix and x is an unknown parameter vector. The random function is observed at t_1, \dots, t_m ,

$$y(t_i) = z(t_i) + n(t_i), \quad i = 1, \dots, m, \quad (22.93)$$

where $n(t_i)$ is the zero-mean random measurement noise. The measurement noise is assumed to be uncorrelated with the signal, $C(s(t), n(t_i)) = 0$, and the covariance matrices of the signal and the measurement noise are assumed to be given as

$$C(s(t_i), s(t_j)) = Q_{s(t_i)s(t_j)} \quad \text{and}$$

$$C(n(t_i), n(t_j)) = Q_{n(t_i)n(t_j)},$$

respectively. If we combine the above equations, we may write in compact vector-matrix form

$$y = Ax + s + n, \quad (22.94)$$

where

$$y = [y(t_1)^\top, \dots, y(t_m)^\top]^\top,$$

$$A = [A_{t_1}^\top, \dots, A_{t_m}^\top]^\top,$$

$$s = [s(t_1)^\top, \dots, s(t_m)^\top]^\top \quad \text{and}$$

$$n = [n(t_1)^\top, \dots, n(t_m)^\top]^\top.$$

We now have all the ingredients available for estimating the trend $A_t x$ and predicting the signal $s(t)$, the noise $n(t_i)$, and the function $z(t)$.

Trend

Let $e = s + n$. Then, $y = Ax + e$ is in the form of a linear model with deterministic parameters, where $E(e) = 0$ and $D(y) = D(e) = Q_{ss} + Q_{nn}$, since $Q_{sn} = 0$. Hence, the BLUE of x is

$$\hat{x} = (A^\top (Q_{ss} + Q_{nn})^{-1} A)^{-1} \times A^\top (Q_{ss} + Q_{nn})^{-1} y. \quad (22.95)$$

Since the trend of the random function is given by its mean, $E(z(t)) = A_t x$, the BLUE of the trend follows as $A_t \hat{x}$.

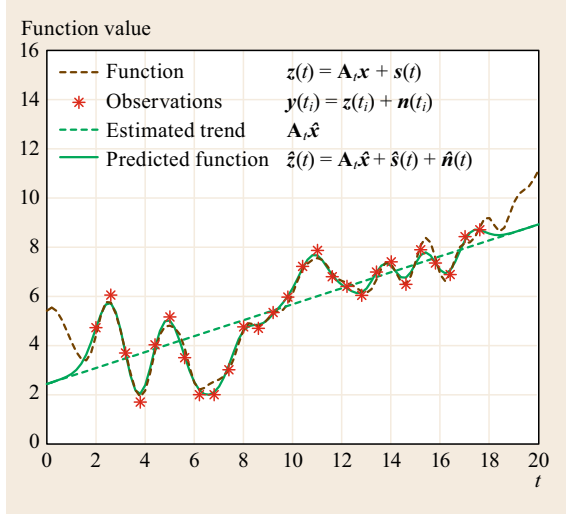


Fig. 22.10 BLUE of the trend $A_t \hat{x}$ and BLUP of the random function $\hat{z}(t)$. Observation values $y(t_i)$, and the random function $z(t)$ are shown as well

Signal

The zero-mean random signal $s(t)$ is the difference between the random function $z(t)$ and its mean $A_t x$. To predict by how much the random function differs from its mean, we need the BLUP of $s(t)$. Since $E(s(t)) = 0$ and $C(s(t), n) = 0$, the BLUP of $s(t)$ is given as

$$\hat{s}(t) = Q_{s(t)s} (Q_{ss} + Q_{nn})^{-1} (y - A\hat{x}). \quad (22.96)$$

Noise

Similar to the prediction of the signal, we can predict the measurement noise $n(t_i)$. We have

$$\hat{n}(t_i) = Q_{n(t_i)n} (Q_{ss} + Q_{nn})^{-1} (y - A\hat{x}). \quad (22.97)$$

Note, since

$$(Q_{s(t_i)s} + Q_{n(t_i)n}) (Q_{ss} + Q_{nn})^{-1} = [0, \dots, \mathbf{I}_k, \dots, 0],$$

that the predicted signal $\hat{s}(t_i)$ and the predicted noise $\hat{n}(t_i)$ add up to the BLUP of $e(t_i) = s(t_i) + n(t_i)$, $\hat{e}(t_i) = \hat{s}(t_i) + \hat{n}(t_i) = y(t_i) - A(t_i)\hat{x}$.

Function

The BLUP of the random function $z(t)$ itself follows as the sum of the BLUE of the trend and the BLUP of the signal

$$\hat{z}(t) = A_t \hat{x} + \hat{s}(t). \quad (22.98)$$

Note, since

$$\hat{s}(t_i) + \hat{n}(t_i) = y(t_i) - A(t_i)\hat{x},$$

that

$$\mathbf{y}(t_i) = \hat{\mathbf{z}}(t_i) + \hat{\mathbf{n}}(t_i) = \mathbf{A}(t_i)\hat{\mathbf{x}} + \hat{\mathbf{s}}(t_i) + \hat{\mathbf{n}}(t_i),$$

which shows that the observable is its own best predictor.

Figure 22.10 shows the function $z(t)$, as well as the BLUE of its trend and its BLUP. In Fig. 22.11, it is shown that the predicted signal $\hat{\mathbf{s}}(t_i)$ and the predicted noise $\hat{\mathbf{n}}(t_i)$ indeed add up to $\hat{\mathbf{e}}(t_i)$. In the figure, this is demonstrated for the BLUP of the random function from Fig. 22.10 at observation instant t_2 .

22.4.3 Properties of MMSE Prediction

The MMSE predictors as discussed in the previous section are all unbiased predictors. Hence, they have a zero-mean prediction error

$$E(\hat{\mathbf{e}}_p) = E(\mathbf{p} - \hat{\mathbf{G}}(\mathbf{y})) = \mathbf{0}. \quad (22.99)$$

The variance matrices of the BLP and BLUP prediction errors are given as [22.21]

$$\begin{aligned} \mathbf{Q}_{\hat{\mathbf{e}}_p \hat{\mathbf{e}}_p}^{\text{BLUP}} &= \mathbf{Q}_{\hat{\mathbf{e}}_p \hat{\mathbf{e}}_p}^{\text{BLP}} + \mathbf{A}_{p|y} \mathbf{Q}_{\hat{\mathbf{x}} \hat{\mathbf{x}}} \mathbf{A}_{p|y}^\top \\ \mathbf{Q}_{\hat{\mathbf{e}}_p \hat{\mathbf{e}}_p}^{\text{BLP}} &= \mathbf{Q}_{pp} - \mathbf{Q}_{py} \mathbf{Q}_{yy}^{-1} \mathbf{Q}_{yp} \end{aligned} \quad (22.100)$$

with

$$\mathbf{A}_{p|y} = \mathbf{A}_p - \mathbf{Q}_{py} \mathbf{Q}_{yy}^{-1} \mathbf{A}_y.$$

Note that

$$\mathbf{Q}_{\hat{\mathbf{e}}_p \hat{\mathbf{e}}_p}^{\text{BLUP}} \geq \mathbf{Q}_{\hat{\mathbf{e}}_p \hat{\mathbf{e}}_p}^{\text{BLP}},$$

22.5 Kalman Filtering

The Kalman filter is a recursive method to estimate the random states of a dynamic system in a way that minimizes the mean squared prediction error. After its initialization, the Kalman filter follows a recursive two-step procedure of time- and measurement-updates. In the time-update (TU), information of the dynamic model is used to predict the system state, and its error-variance matrix, ahead in time, thus, giving $\hat{\mathbf{x}}_{k|k-1}$ and its error variance matrix $\mathbf{P}_{k|k-1}$. In the measurement-update (MU), the newly arrived measurements \mathbf{y}_k are combined, in a MMSE-sense, with the predicted state $\hat{\mathbf{x}}_{k|k-1}$ to obtain the filtered state $\hat{\mathbf{x}}_{k|k}$ and its error variance matrix $\mathbf{P}_{k|k}$ (Fig. 22.12).

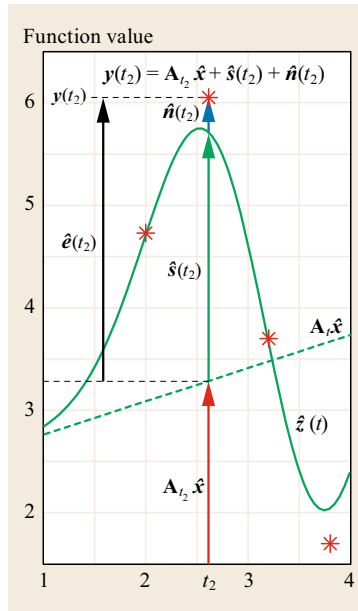


Fig. 22.11
Portion of Fig. 22.10, showing that $\hat{\mathbf{e}}(t_2) = \hat{\mathbf{s}}(t_2) + \hat{\mathbf{n}}(t_2) = \mathbf{y}(t_2) - \mathbf{A}_{t_2} \hat{\mathbf{x}}$

that is, the BLUP prediction error is never more precise than that of the BLP. This is due to additional uncertainty that enters by having to estimate \mathbf{x} as well.

The variance-covariance matrix of the prediction error, also referred to as *error variance matrix*, should not be confused with the variance-covariance matrix of the predictor itself, which in case of BLUP is given as

$$\mathbf{Q}_{\hat{\mathbf{p}} \hat{\mathbf{p}}}^{\text{BLUP}} = \mathbf{Q}_{\hat{\mathbf{p}} \hat{\mathbf{p}}}^{\text{BLUE}} + \mathbf{Q}_{py} \mathbf{Q}_{yy}^{-1} \mathbf{Q}_{ee} \mathbf{Q}_{yy}^{-1} \mathbf{Q}_{yp} \quad (22.101)$$

with

$$\mathbf{Q}_{\hat{\mathbf{p}} \hat{\mathbf{p}}}^{\text{BLUE}} = \mathbf{A}_p \mathbf{Q}_{\hat{\mathbf{x}} \hat{\mathbf{x}}} \mathbf{A}_p^\top.$$

22.5.1 Model Assumptions

First, we state the assumptions on which the Kalman filter is based. They concern the measurement model and the dynamic model.

The Measurement Model

The link between the random vector of observables \mathbf{y}_i and the random state-vector \mathbf{x}_i is assumed given as

$$\mathbf{y}_i = \mathbf{A}_i \mathbf{x}_i + \mathbf{n}_i, \quad i = 0, 1, \dots, t \quad (22.102)$$

together with

$$E(\mathbf{x}_0) = \bar{\mathbf{x}}_0, \quad E(\mathbf{n}_i) = \mathbf{0} \quad (22.103)$$

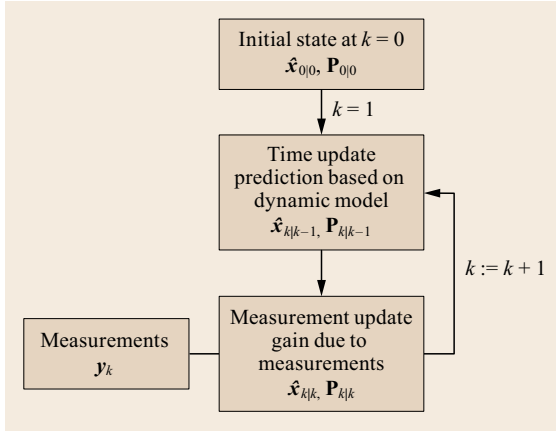


Fig. 22.12 Kalman filter recursion with time-update (TU) and measurement-update (MU)

and

$$\mathbf{C}(\mathbf{x}_0, \mathbf{n}_i) = \mathbf{0}, \quad \mathbf{C}(\mathbf{n}_i, \mathbf{n}_j) = \mathbf{R}_i \delta_{ij}, \quad i = 0, 1, \dots, t, \quad (22.104)$$

where $\mathbf{C}(\mathbf{u}, \mathbf{v})$ denotes the covariance between two random vectors \mathbf{u} and \mathbf{v} and δ_{ij} is the Kronecker delta. Thus, the zero-mean measurement noise \mathbf{n}_i is assumed to be uncorrelated in time and to be uncorrelated with the initial state-vector \mathbf{x}_0 .

The Dynamic Model

The linear dynamic model, describing the time-evolution of the random state-vector \mathbf{x}_i , is given as

$$\mathbf{x}_i = \Phi_{i,i-1} \mathbf{x}_{i-1} + \mathbf{d}_i, \quad i = 1, 2, \dots, t \quad (22.105)$$

with

$$E(\mathbf{d}_i) = \mathbf{0}, \quad \mathbf{C}(\mathbf{x}_0, \mathbf{d}_i) = \mathbf{0}, \quad (22.106)$$

$$\mathbf{C}(\mathbf{d}_i, \mathbf{n}_j) = \mathbf{0}, \quad \mathbf{C}(\mathbf{d}_i, \mathbf{d}_j) = \mathbf{S}_i \delta_{ij}, \quad i, j = 1, 2, \dots, t, \quad (22.107)$$

where $\Phi_{i,i-1}$ denotes the transition matrix and the random vector \mathbf{d}_i is the system noise. The system noise \mathbf{d}_i is thus also assumed to have a zero mean, to be uncorrelated in time and to be uncorrelated with the initial state-vector and the measurement noise. Note that the transition matrix from epoch j to i is denoted as Φ_{ij} . Thus, $\Phi_{i,j}^{-1} = \Phi_{j,i}$.

The dynamic model (22.105) and its transition matrix can be obtained from solving the dynamic system's first-order vectorial differential equation. For instance, if the dynamic system is described by a first-order linearized time-varying system of differential equations,

$\dot{\mathbf{x}}_t = \mathbf{F}_t \mathbf{x}_t + \mathbf{G}_t \mathbf{u}_t$, then

$$\mathbf{x}_t = \Phi_{t,t_0} \mathbf{x}_{t_0} + \int_{t_0}^t \Phi_{t,\tau} \mathbf{G}_\tau \mathbf{u}_\tau d\tau \quad (22.108)$$

with the transition matrix being the solution of [22.24, 25]

$$\frac{\partial \Phi_{t,t_0}}{\partial t} = \mathbf{F}_t \Phi_{t,t_0}, \quad \Phi_{t,t_0} = \mathbf{I}_n.$$

Methods for its numerical integration can be found in, for example, [22.26]. For the case matrix \mathbf{F}_t is time-invariant, the solution is given by the matrix exponential $\Phi_{t,t_0} = \exp(\mathbf{F}(t-t_0))$. Methods of evaluating the matrix exponential are given in [22.27].

22.5.2 The Kalman Filter Recursion

The Kalman filter is usually derived as either a recursive BP (Gaussian case) or as a recursive BLP, [22.28–31]. Both these predictors require that the mean of the random state vector, $E(\mathbf{x}_i)$, is known. This is why in the derivation of the Kalman filter one usually assumes the mean of the random initial state-vector $\bar{\mathbf{x}}_0$ (22.103) to be known [22.32–35]. Such derivation of the Kalman filter is however not appropriate in case the mean of the random state vector is unknown, a situation that applies to most engineering applications. In the following, we will therefore assume $E(\mathbf{x}_0) = \bar{\mathbf{x}}_0$ unknown and present the results obtained when applying the recursive BLUP principle. We start with the initialization and then present the TU and MU.

Initialization

The initial state and its error-variance matrix are given as

$$\begin{aligned} \hat{\mathbf{x}}_{0|0} &= (\mathbf{A}^\top \mathbf{R}_0^{-1} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{R}_0^{-1} \mathbf{y}_0 \\ \mathbf{P}_{0|0} &= (\mathbf{A}^\top \mathbf{R}_0^{-1} \mathbf{A})^{-1}. \end{aligned} \quad (22.109)$$

Note that $\mathbf{P}_{0|0}$ is *not* the variance–covariance matrix of $\hat{\mathbf{x}}_{0|0}$. If $\mathbf{Q}_{\mathbf{x}_0 \mathbf{x}_0}$ is the variance–covariance matrix of the random state vector \mathbf{x}_0 , then the sum $\mathbf{Q}_{\mathbf{x}_0 \mathbf{x}_0} + \mathbf{P}_{0|0}$ is the variance–covariance matrix of $\hat{\mathbf{x}}_{0|0}$. Would one assume that the mean of the random state vector is known, then the initialization would be given as $\hat{\mathbf{x}}_{0|0} = \bar{\mathbf{x}}_0$, with error-variance matrix $\mathbf{P}_{0|0} = \mathbf{Q}_{\mathbf{x}_0 \mathbf{x}_0}$.

Time Update (TU)

The time-updated state vector and its error-variance matrix are given as

$$\begin{aligned} \hat{\mathbf{x}}_{k|k-1} &= \Phi_{k,k-1} \hat{\mathbf{x}}_{k-1|k-1} \\ \mathbf{P}_{k|k-1} &= \Phi_{k,k-1} \mathbf{P}_{k-1|k-1} \Phi_{k,k-1}^\top + \mathbf{S}_k, \end{aligned} \quad (22.110)$$

for $k = 1, \dots, t$.

Measurement-Update (MU)

In the MU, it is the predicted residual that is used to correct the predicted state. The predicted residual (sometimes also referred to as *innovation*) and its variance–covariance matrix are given as

$$\begin{aligned} \mathbf{v}_k &= \mathbf{y}_k - \mathbf{A}_k \hat{\mathbf{x}}_{k|k-1} \\ \mathbf{Q}_{\mathbf{v}_k \mathbf{v}_k} &= \mathbf{R}_k + \mathbf{A}_k \mathbf{P}_{k|k-1} \mathbf{A}_k^\top. \end{aligned} \quad (22.111)$$

The predicted residuals have the important property that they are zero-mean and uncorrelated in time

$$E(\mathbf{v}_k) = \mathbf{0}, \quad \mathbf{C}(\mathbf{v}_k, \mathbf{v}_l) = \mathbf{Q}_{\mathbf{v}_k \mathbf{v}_k} \delta_{k,l}. \quad (22.112)$$

This property forms the basis for performing recursive quality control and hypotheses testing (Chap. 24).

Using the predicted residual vector, the filtered state and its error-variance matrix are given as

$$\begin{aligned} \hat{\mathbf{x}}_{k|k} &= \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \mathbf{v}_k \\ \mathbf{P}_{k|k} &= (\mathbf{I}_n - \mathbf{K}_k \mathbf{A}_k) \mathbf{P}_{k|k-1} \end{aligned} \quad (22.113)$$

with the Kalman gain matrix given as

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{A}_k^\top \mathbf{Q}_{\mathbf{v}_k \mathbf{v}_k}^{-1}. \quad (22.114)$$

An illustration of the contributions of the time update (22.110) and the measurement update (22.113) is shown in Fig. 22.13. The green line shows the actual trajectory \mathbf{x}_t , not to be confused with the mean trajectory $E(\mathbf{x}_t) = \bar{\mathbf{x}}_t$. Instead of recovering the actual trajectory, as is done recursively with the Kalman filter, an added objective can be to estimate the unknown mean trajectory $\bar{\mathbf{x}}_t$ as well. Also this can be done recursively, thereby making use of the Kalman filter outputs. The corresponding BLUE–BLUP recursion for the recovery of the mean trajectory is given in [22.36]. In the absence of system noise ($\mathbf{S}_k = \mathbf{0}, k = 1, \dots$), the two solutions, that is, for the mean trajectory and for the actual trajectory, coincide. And, if additionally $\Phi_{i,i-1} = \mathbf{I}$, then they reduce further to the recursive solution (22.38).

Example 22.11 Ionospheric Delays

Figure 22.14 presents an example of the filtered solution of the ionospheric delays for a certain satellite; it is based on real data. The observations (shown in gray) are given by a time series of the ionosphere combination of the GPS code observations p_1 and p_2 on frequencies L1 and L2, respectively

$$\iota_p(t) = \begin{bmatrix} -\frac{f_2^2}{f_1^2 - f_2^2} & \frac{f_2^2}{f_1^2 - f_2^2} \end{bmatrix} \begin{bmatrix} p_1(t) \\ p_2(t) \end{bmatrix}. \quad (22.115)$$

It is assumed that second-order derivative of ionospheric delay is zero mean constant, and the transition matrix

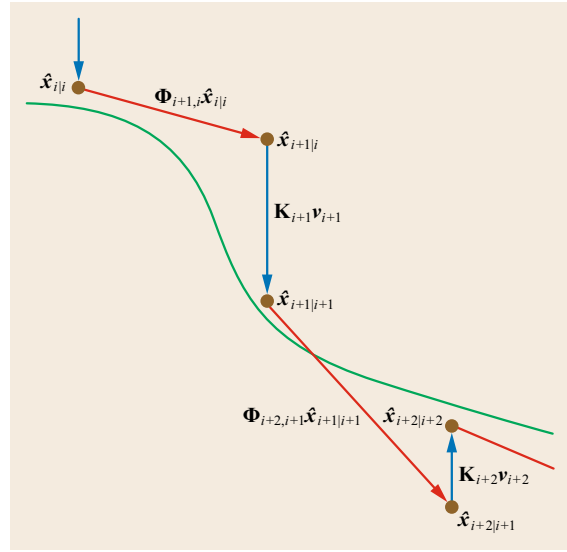


Fig. 22.13 Contributions of time update and measurement update: actual trajectory \mathbf{x}_t is shown in green. The predicted states $\hat{\mathbf{x}}_{k|k-1}$ based on the time update (red arrows), as well as the filtered states $\hat{\mathbf{x}}_{k|k}$ after the measurement update (blue arrows) are shown

$\Phi_{k,k-1}$ and system noise \mathbf{S}_k are therefore given as

$$\Phi_{k,k-1} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}, \quad \mathbf{S}_k = \sigma_d^2 \begin{bmatrix} \frac{1}{2} \Delta t^4 & \frac{1}{2} \Delta t^3 \\ \frac{1}{2} \Delta t^3 & \Delta t^2 \end{bmatrix}. \quad (22.116)$$

In this case, Δt is 30 s, and σ_d^2 is the variance factor of the process noise.

Both the predicted and filtered states as well as their difference are shown in Fig. 22.14. In order to show the effect of the system noise, σ_d^2 was reduced by a factor 100, and the corresponding results are shown in the bottom panel. Obviously, the contribution of the observations becomes much smaller, and consequently the difference between predicted and filtered states is smaller as well. The filtered solution is now much smoother, but may deviate considerably from the observed delays.

22.5.3 Kalman Filter Information Form

The measurement update equations (22.113) are given in the *variance form*, since they are expressed in terms of the variance–covariance matrices \mathbf{R}_k and $\mathbf{P}_{k|k-1}$. An alternative, the so-called *information form*, is also possible. In the information form, the inverses of \mathbf{R}_k and $\mathbf{P}_{k|k-1}$ are used. The Kalman gain matrix \mathbf{K}_k and error-variance matrix $\mathbf{P}_{k|k}$ can be expressed in \mathbf{R}_k^{-1} and $\mathbf{P}_{k|k-1}^{-1}$

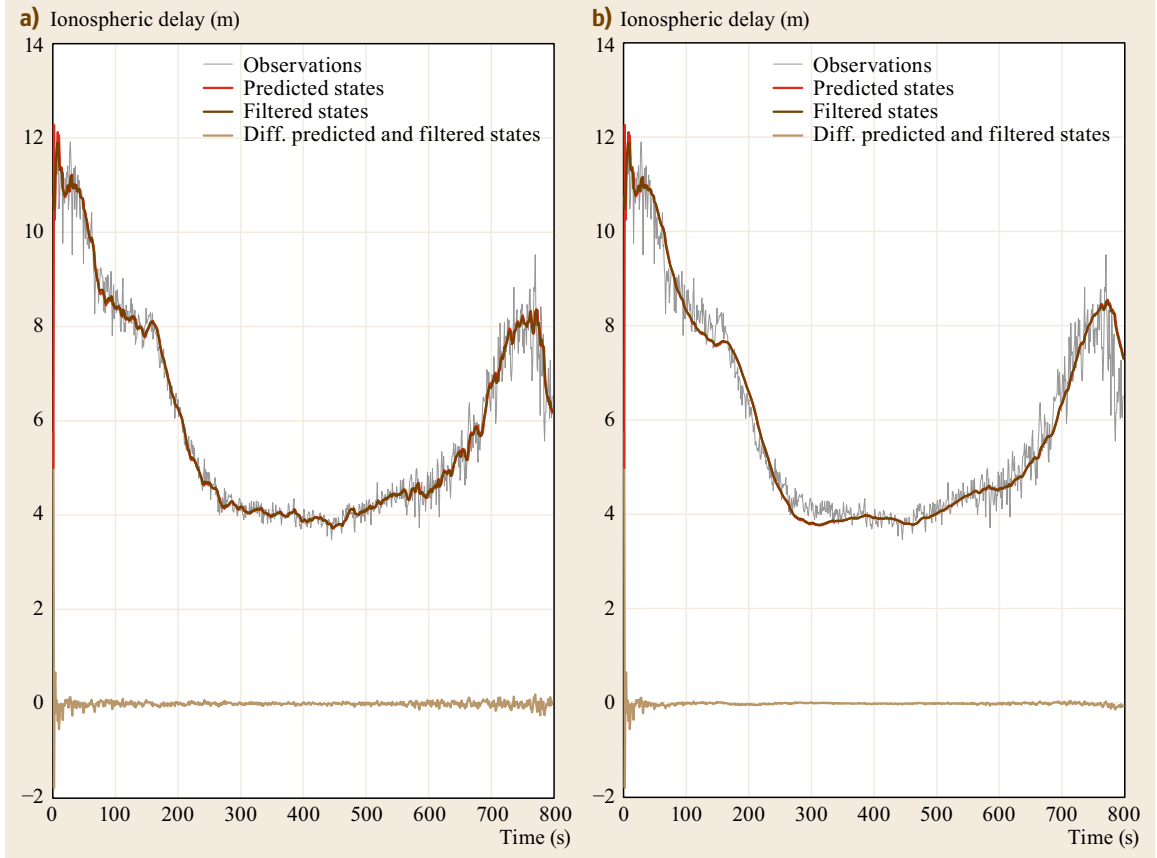


Fig. 22.14a,b Observations of ionospheric delays based on ionosphere combination of GPS L1 and L2 code observations, as well as predicted and filtered states from Kalman filtering results. It is assumed that the second-order derivative of the ionospheric delays is a random constant process with zero-mean. **(a)** Realistic σ_d^2 ; **(b)** $\sigma_d^2/100$

as

$$\begin{aligned} \mathbf{K}_k &= \mathbf{P}_{k|k} \mathbf{A}_k^\top \mathbf{R}_k^{-1} \\ \mathbf{P}_{k|k} &= (\mathbf{P}_{k|k-1}^{-1} + \mathbf{A}_k^\top \mathbf{R}_k^{-1} \mathbf{A}_k)^{-1}. \end{aligned} \quad (22.117)$$

Although the variance form and information form both give identical results, one form may be preferred over the other if one considers the task of matrix inversion. In general, the variance form is preferred over the information form if the dimension of \mathbf{y}_k is (much) smaller than that of \mathbf{x}_k . For instance, in case \mathbf{y}_k is a scalar, the variance-form would need the trivial inversion of the scalar $\mathbf{Q}_{v_k v_k}$, while the information form would then still need the inversion of $\mathbf{P}_{k|k-1}^{-1}$.

The square root information filter (SRIF) is an efficient mechanization of the Kalman filter [22.37, 38]. Similarly, as in Sect. 22.1.3, it uses a decomposition of the variance-covariance matrices in square root form.

22.5.4 Extended Kalman Filter

The Kalman filter applies to models that are linear. In practice, however, measurement and/or dynamic models are often nonlinear. One would then have nonlinear vector functions $\mathbf{A}_k(\cdot)$ and $\Phi_k(\cdot)$, instead of the matrices \mathbf{A}_k and $\Phi_{k,k-1}$ of (22.102) and (22.105), respectively. The extended Kalman filter (EKF) is the Kalman filter applied to the *linearized* versions of such nonlinear measurement and dynamic models [22.29, 35, 39]. Thus, if the nonlinear equations are given as

$$\begin{aligned} \mathbf{y}_k &= \mathbf{A}_k(\mathbf{x}_k) + \mathbf{n}_k \\ \mathbf{x}_k &= \Phi_k(\mathbf{x}_{k-1}) + \mathbf{d}_k, \end{aligned} \quad (22.118)$$

the EKF TU is given by

$$\begin{aligned} \hat{\mathbf{x}}_{k|k-1} &= \Phi_k(\hat{\mathbf{x}}_{k-1|k-1}) \\ \mathbf{P}_{k|k-1} &= \mathbf{J}_{\Phi_k} \mathbf{P}_{k-1|k-1} \mathbf{J}_{\Phi_k}^\top + \mathbf{S}_k \end{aligned} \quad (22.119)$$

and the EKF MU by

$$\begin{aligned}\hat{\mathbf{x}}_{k|k} &= \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k(\mathbf{y}_k - \mathbf{A}_k(\hat{\mathbf{x}}_{k|k-1})) \\ \mathbf{P}_{k|k} &= (\mathbf{I}_n - \mathbf{K}_k \mathbf{J}_{\mathbf{A}_k}) \mathbf{P}_{k|k-1}\end{aligned}\quad (22.120)$$

with the gain matrix given as

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{J}_{\mathbf{A}_k}^\top (\mathbf{R}_k + \mathbf{J}_{\mathbf{A}_k} \mathbf{P}_{k|k-1} \mathbf{J}_{\mathbf{A}_k}^\top)^{-1}, \quad (22.121)$$

where $\mathbf{J}_{\mathbf{A}_k}$ is the Jacobian of $\mathbf{A}_k(\cdot)$ evaluated at $\hat{\mathbf{x}}_{k|k-1}$ and \mathbf{J}_{Φ_k} is the Jacobian of $\Phi_k(\cdot)$ evaluated at $\hat{\mathbf{x}}_{k-1|k-1}$. If $\mathbf{A}_k(\cdot)$ is highly nonlinear, a further improvement can be obtained by iterating the solution of the EKF measurement update. This is referred to as the iterated extended Kalman filter (IEKF). As the IEKF can be shown to be a special case of a Gauss–Newton iteration, it also inherits all its convergency properties [22.40].

22.5.5 Smoothing

Smoothing filters rely on forward running Kalman filters. They are often used in offline (post-) processing in order to get better estimates of the past states based on all available information, but can also be used parallel to a filter running in realtime. Three types of smoothers are distinguished as follows, where N refers to the length of the interval and k is the instant at which a smoothed estimate is needed.

The first is the *fixed-interval* smoother, also known as the Rauch–Tung–Striebel filter or forward–backward filter. The smoothing interval is fixed to a length N , but the time instant t_k for which the smoothed estimate is computed changes. It is based on the Kalman filter solutions up till and including epoch $k < N$, and a backward running filter from epoch N to k .

The *fixed-point* smoother is estimating the state at a fixed time instant t_k , and this state estimate is updated as more and more observations become available, that is, N is increasing.

Finally, the *fixed-lag* smoother is estimating the states with a fixed lag of N epochs using the observations up till and including current epoch. Hence, the final smoothed state for a given time instant t_k is computed after N epochs and $k+N$ is the last observation epoch that is used.

Figures 22.15–22.17 illustrate the different principles.

Fixed-Interval Smoothing

The Rauch–Tung–Striebel smoothing algorithm was proposed in [22.41], and is the fixed-interval smoother that will be discussed here. For each t_k in the interval $[t_0, t_N]$, the state estimate is based on all observations in this interval. The idea is illustrated in Fig. 22.15: The forward filter is the standard Kalman filter to provide the state estimate at an epoch k based on epochs 0 till k ;

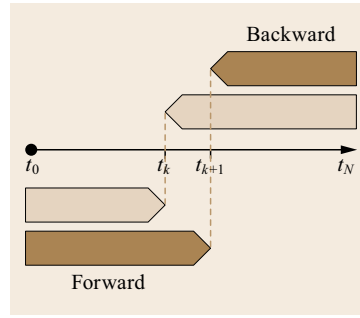


Fig. 22.15 Fixed-interval smoothing based on forward–backward filtering to estimate the state at any time t_i ($i = 0, \dots, N$) using all available observations from t_0 till t_N . The principle is shown here for times t_k and t_{k+1}

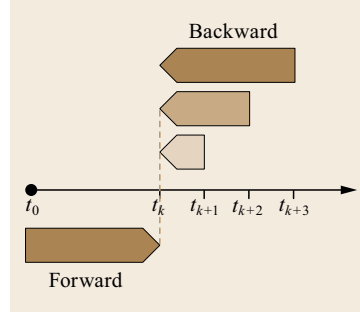


Fig. 22.16 Fixed-point smoothing to estimate the state at fixed time t_k with increasing smoothing window as more observations become available

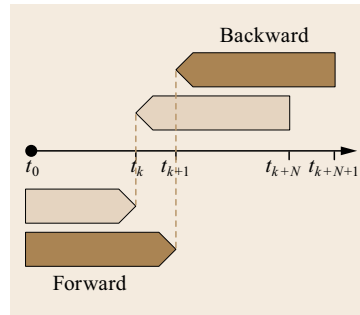


Fig. 22.17 Fixed-lag smoothing to estimate the state at any time t_i using all available observations from t_0 till t_{i+N} , that is, with a smoothing interval of fixed length N . The principle is shown here for times t_k and t_{k+1}

the backward filter is applied to obtain a smoothed state using the observations from k till N .

From the forward filter, we have the estimated states $\hat{\mathbf{x}}_{k|k-1}$ and $\hat{\mathbf{x}}_{k|k}$, together with their error variance–covariance matrices $\mathbf{P}_{k|k-1}$ and $\mathbf{P}_{k|k}$.

The backward filter runs recursively with $k = N-1, N-2, \dots, 0$. The smoothed state estimate and its error variance–covariance matrix equal

$$\hat{\mathbf{x}}_{k|N} = \hat{\mathbf{x}}_{k|k} + \mathbf{L}_k (\hat{\mathbf{x}}_{k+1|N} - \hat{\mathbf{x}}_{k+1|k}) \quad (22.122)$$

$$\mathbf{P}_{k|N} = \mathbf{P}_{k|k} + \mathbf{L}_k (\mathbf{P}_{k+1|N} - \mathbf{P}_{k+1|k}) \mathbf{L}_k^\top, \quad (22.123)$$

where

$$\mathbf{L}_k = \mathbf{P}_{k|k} \Phi_{k+1,k}^\top \mathbf{P}_{k+1|k}^{-1}. \quad (22.124)$$

The fixed-interval smoother is mainly used for offline processing, as storage of all results from the forward filter is required and inversion of $\mathbf{P}_{k+1|k}$ is needed for every recursion.

Example 22.12 Ionospheric delays (continued)

For the example of the filtered ionosphere delay solution, presented at the end of Sect. 22.5.2, also the fixed-interval smoother was applied. The complete observation interval is used for smoothing the state at each epoch. Figure 22.18 shows the smoothed states. For comparison, also the filtered states are shown. The effect of the forward–backward filtering is obvious for the whole period, and very pronounced at the start of the observation interval, since the Kalman filter needs to converge.

Fixed-Point Smoothing

With fixed-point smoothing [22.42], the state estimate for one specific t_k is continuously smoothed as new observations become available, that is, the time win-

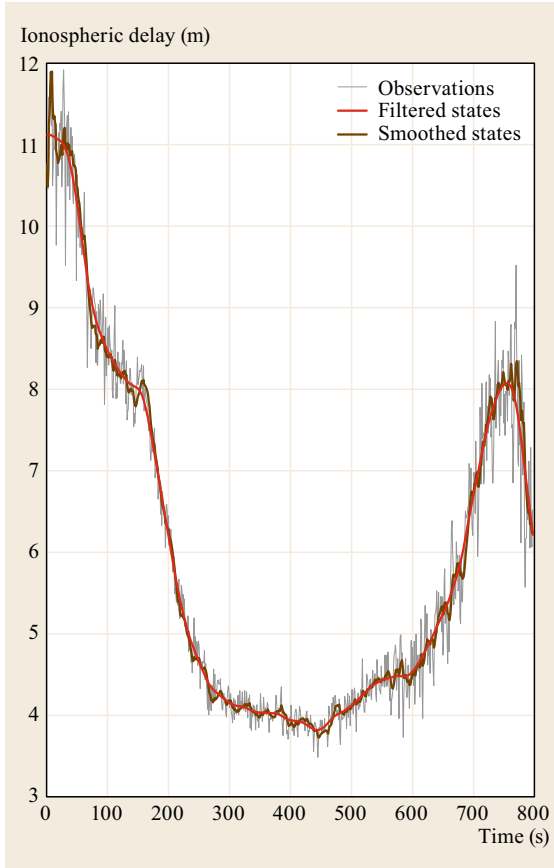


Fig. 22.18 Observations of ionospheric delays based on ionosphere combination of GPS L1 and L2 code observations, as well as filtered and smoothed states from Kalman filtering and fixed-interval smoothing, respectively. It is assumed that the second-order derivative of the ionospheric delays is a random constant process with zero mean

dow used to estimate the smoothed state becomes much longer. In the following, the superscript s indicates the smoothed state; estimates and variance–covariance matrices without the superscript refer to the standard Kalman filter solutions.

As initial values, the standard Kalman filter solution is used: $\hat{\mathbf{x}}_{k|k}^s = \hat{\mathbf{x}}_{k|k}$, $\mathbf{P}_{k|k}^s = \mathbf{P}_{k|k}$. Then, for $j = k+1, k+2, \dots$ (and k fixed)

$$\hat{\mathbf{x}}_{k|j}^s = \hat{\mathbf{x}}_{k|j-1}^s + \mathbf{M}_j(\hat{\mathbf{x}}_{j|j} - \hat{\mathbf{x}}_{j|j-1}), \quad (22.125)$$

$$\mathbf{P}_{k|j}^s = \mathbf{P}_{k|j-1}^s + \mathbf{M}_j(\mathbf{P}_{j|j} - \mathbf{P}_{j|j-1})\mathbf{M}_j^\top, \quad (22.126)$$

where

$$\mathbf{M}_j = \prod_{i=k}^{j-1} \mathbf{L}_i \quad (22.127)$$

with \mathbf{L}_i from (22.124).

Fixed-Lag Smoothing

Fixed-lag smoothing [22.42] implies that the final smoothed state of a certain epoch is calculated after a fixed lag of N epochs. The smoothed states are obtained by a recursion over $k = 0, 1, 2, \dots$

$$\hat{\mathbf{x}}_{k|k+N}^s = \mathbf{p}_1 + \mathbf{p}_2 + \mathbf{p}_3, \quad (22.128)$$

with

$$\mathbf{p}_1 = \Phi_{k,k-1} \hat{\mathbf{x}}_{k-1|k-1+N}^s$$

$$\mathbf{p}_2 = \mathbf{S}_{k-1} \Phi_{k,k-1}^{-\top} \mathbf{P}_{k-1|k-1}^{-1} (\hat{\mathbf{x}}_{k-1|k-1+N}^s - \hat{\mathbf{x}}_{k-1|k-1})$$

$$\mathbf{p}_3 = \mathbf{M}_{k+N} \mathbf{K}_{k+N} \mathbf{v}_{k+N}.$$

The terms \mathbf{p}_1 and \mathbf{p}_2 form the prediction based on the smoothed state at epoch $k-1$; \mathbf{p}_3 is the contribution of the new observations at epoch $k+N$.

The error variance–covariance matrix of the smoothed state equals

$$\mathbf{P}_{k|k+N}^s = \mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3, \quad (22.129)$$

with

$$\mathbf{q}_1 = \mathbf{P}_{k|k-1}^s$$

$$\mathbf{q}_2 = -\mathbf{L}_{k-1}^{-1} (\mathbf{P}_{k-1|k-1} - \mathbf{P}_{k-1|k-1+N}^s) \mathbf{L}_{k-1}^{-\top}$$

$$\mathbf{q}_3 = \mathbf{M}_{k+N} \mathbf{K}_{k+N} \mathbf{A}_{k+N} \mathbf{P}_{k+N|k-1+N} \mathbf{M}_{k+N}^\top.$$

As can be seen from (22.128) and (22.129), the input from a running Kalman filter is required.

The initial conditions $\hat{\mathbf{x}}_{0|N}^s$ and $\mathbf{P}_{0|N}^s$ must be determined with the fixed-point smoother.

Acknowledgments. The second author is the recipient of an Australian Research Council Federation Fellowship (project number FF0883188). This support is gratefully acknowledged.

References

- 22.1 H.W. Sorenson: Least-squares estimation: From Gauss to Kalman, *IEEE Spectr.* **7**(7), 63–68 (1970)
- 22.2 P.J.G. Teunissen: *Adjustment Theory, an Introduction* (Delft Academic, Delft 2004)
- 22.3 G.H. Golub, C.F. van Loan: *Matrix Computations*, Vol. 3 (Johns Hopkins Univ. Press, Baltimore 2012)
- 22.4 J.M. Ortega: *Matrix Theory: A Second Course* (Plenum, New York 1987)
- 22.5 C.L. Lawson, R.J. Hanson: *Solving Least-Squares Problems* (Prentice-Hall, Eaglewood Cliffs 1974)
- 22.6 K.R. Koch: *Parameter Estimation and Hypothesis Testing in Linear Models* (Springer, Berlin 1999)
- 22.7 R. Hatch: The synergism of GPS code and carrier measurements, *Proc. 3rd Int. Geod. Symp. Satell. Doppler Position*. (Physical Science Laboratory, Las Cruces 1982) pp. 1213–1232
- 22.8 R. Hatch: Dynamic differential GPS at the centimeter level, *Proc. 4th Int. Geod. Symp. Satell. Position.*, Austin (Defense Mapping Agency/National Geodetic Survey, Silver Spring 1986) pp. 1287–1298
- 22.9 P.J.G. Teunissen: The GPS phase-adjusted pseudorange, *Proc. 2nd Int. Workshop High Precis. Navig.*, ed. by K. Linkwitz, U. Hangleiter (Dümmler, Bonn 1991) pp. 115–125
- 22.10 H. Wolf: The Helmert block method, its origin and development, *Proc. 2nd Int. Symp. Probl. Relat. Redefinition North Am. Geod. Netw.*, Arlington (U.S. Dept. of Commerce, Virginia 1978) pp. 319–326
- 22.11 H. Boomkamp: Distributed processing for large geodetic solutions, *Proc. Int. Assoc. Geod. Symp.*, Marne-La-Vallee, ed. by Z. Altamimi, X. Collilieux (Springer, Berlin Heidelberg 2013) pp. 13–18
- 22.12 P. Davies, G. Blewitt: Methodology for global geodetic time series estimation: A new tool for geodynamics, *J. Geophys. Res.* **105**(B5), 11083–11100 (2000)
- 22.13 A.A. Lange: Fast Kalman processing of the GPS carrier-phases for mobile positioning and atmospheric tomography, *Proc. FIG Work. Week Surv. Key Role Accel. Dev. Eilat* (2009)
- 22.14 D.G. Pursell, M. Potterfield: *National Readjustment Final Report (NOAA, Silver Spring 2008), NOAA Technical Report NOS, NAD 83 (NSRS, Vol. 60, 2007)*
- 22.15 P.J.G. Teunissen: *Network Quality Control* (Delft Academic, Delft 2006)
- 22.16 D. Odijk, B. Zhang, A. Khodabandeh, P.J.G. Teunissen: On the estimability of parameters in undifferenced, uncombined GNSS network and PPP-RTK user models by means of S-system theory, *J. Geod.* **90**(1), 15–44 (2016)
- 22.17 P.J.G. Teunissen: Quality control in geodetic networks. In: *Optimization and Design of Geodetic Networks*, ed. by E.W. Grafarend, F. Sansò (Springer, Berlin 1985) pp. 526–547
- 22.18 P.J.G. Teunissen: Nonlinear least-squares, *Manuscr. Geodaetica* **15**(3), 137–150 (1990)
- 22.19 A. Rusczyński: *Nonlinear Optimization* (Princeton Univ. Press, Princeton 2006)
- 22.20 P.J.G. Teunissen: First and second moments of nonlinear least-squares estimators, *J. Geod.* **63**(3), 253–262 (1989)
- 22.21 P.J.G. Teunissen: Best prediction in linear models with mixed integer/real unknowns: Theory and application, *J. Geod.* **81**(12), 759–780 (2007)
- 22.22 C. Alber, R. Ware, C. Rocken, J. Braun: Inverting GPS double differences to obtain single path phase delays, *Geophys. Res. Lett.* **27**, 2661–2664 (2000)
- 22.23 H. van der Marel, B. Gundlich: *Development of Models for Use of Slant Delays, Slant Delay Retrieval and Multipath Mapping Software* (Danish Meteorological Institute, Copenhagen 2006)
- 22.24 R.A. Decarlo: *Linear Systems, a State Variable Approach with Numerical Implementation* (Prentice-Hall, Upper Saddle River 1989)
- 22.25 P.J.G. Teunissen: *Dynamic Data Processing* (Delft Academic, Delft 2001)
- 22.26 J. Stoer, R. Bulirsch: *Introduction to Numerical Analysis*, 2nd edn. (Springer, New York 1994)
- 22.27 N.J. Higham: *Functions of Matrices: Theory and Computation* (Society for Industrial and Applied Mathematics, Philadelphia 2009)
- 22.28 R.E. Kalman: A new approach to linear filtering and prediction problems, *J. Basic Eng.* **82**(1), 35–45 (1960)
- 22.29 A. Gelb: *Applied Optimal Estimation* (MIT Press, Cambridge 1974)
- 22.30 A.H. Jazwinski: *Stochastic Processes and Filtering Theory* (Dover Publications, Mineola 1991)
- 22.31 T. Kailath: *Lectures on Wiener and Kalman Filtering* (Springer, Berlin 1981)
- 22.32 P.S. Maybeck: *Stochastic Models, Estimation, and Control* (Academic, New York 1979)
- 22.33 B.D.O. Anderson, J.B. Moore: *Optimal Filtering* (Prentice-Hall, Englewood Cliffs 1979)
- 22.34 H. Stark, J. Woods: *Probability, Random Processes, and Estimation Theory for Engineers* (Prentice-Hall, Englewood Cliffs 1986)
- 22.35 D. Simon: *Optimal State Estimation: Kalman, H_∞ and Nonlinear Approaches* (Wiley, New York 2006)
- 22.36 A. Khodabandeh, P.J.G. Teunissen: A recursive linear MMSE filter for dynamic systems with unknown state vector means, *GEM – Int. J. Geomath.* **5**(1), 17–31 (2014)
- 22.37 G.J. Bierman: *Factorization Methods for Discrete Sequential Estimation* (Academic, New York 1977)
- 22.38 C.C.J.M. Tiberius: *Recursive Data Processing for Kinematic GPS Surveying* (Publications on Geodesy, 45, Netherlands Geodetic Commission, Delft 1998)
- 22.39 R.G. Brown, P.Y.C. Hwang: *Introduction to Random Signals and Applied Kalman Filtering: With MATLAB Exercises and Solutions* (Wiley, New York 2012)
- 22.40 P.J.G. Teunissen: On the local convergence of the iterated extended Kalman filter, *Proc. XX Gen. Assembly IUGG, IAG Section IV, Vienna* (1991) pp. 177–184

- 22.41 H.E. Rauch, F. Tung, C.T. Striebel: Maximum likelihood estimates of linear dynamic systems, AIAA J. 3(8), 1445–1450 (1965)
- 22.42 J.S. Meditch: *Stochastic Optimal Linear Estimation and Control* (McGraw-Hill, New York 1969)

23. Carrier Phase Integer Ambiguity Resolution

Peter J.G. Teunissen

Global Navigation Satellite System (GNSS) carrier-phase integer ambiguity resolution is the process of resolving the carrier-phase ambiguities as integers. It is the key to fast and high-precision GNSS parameter estimation and it applies to a great variety of GNSS models that are currently in use in navigation, surveying, geodesy and geophysics. The theory that underpins GNSS carrier-phase ambiguity resolution is the theory of integer inference. This theory and its practical application is the topic of the present chapter.

23.1	GNSS Ambiguity Resolution	662
23.1.1	The GNSS Model.....	662
23.1.2	Ambiguity Resolution Steps.....	662
23.1.3	Ambiguity Resolution Quality.....	663
23.2	Rounding and Bootstrapping	666
23.2.1	Integer Rounding.....	666
23.2.2	Vectorial Rounding.....	666

23.2.3	Integer Bootstrapping.....	667
23.2.4	Bootstrapped Success Rate.....	668
23.3	Linear Combinations	669
23.3.1	Z-transformations.....	669
23.3.2	(Extra) Widening.....	669
23.3.3	Decorrelating Transformation.....	670
23.3.4	Numerical Example.....	672
23.4	Integer Least-Squares	673
23.4.1	Mixed Integer Least-Squares.....	673
23.4.2	The ILS Computation.....	674
23.4.3	Least-Squares Success Rate.....	676
23.5	Partial Ambiguity Resolution	677
23.6	When to Accept the Integer Solution?	678
23.6.1	Model- and Data-Driven Rules.....	678
23.6.2	Four Ambiguity Resolution Steps.....	679
23.6.3	Quality of Accepted Integer Solution.....	679
23.6.4	Fixed Failure-Rate Ratio Test.....	680
23.6.5	Optimal Integer Ambiguity Test.....	681
	References	683

Carrier-phase integer ambiguity resolution is the key to fast and high-precision GNSS parameter estimation. It is the process of resolving the unknown cycle ambiguities of the carrier-phase data as integers. Once this has been done successfully, the very precise carrier-phase data will act as very precise pseudorange data, thus making very precise positioning and navigation possible.

GNSS ambiguity resolution applies to a great variety of current and future GNSS models, with applications in surveying, navigation, geodesy and geophysics. These models may differ greatly in complexity and diversity. They range from single-receiver or single-baseline models used for kinematic positioning to multibaseline models used as a tool for studying geodynamic phenomena. The models may or may not have the relative receiver-satellite geometry included. They may also be discriminated as to whether the slave receiver(s) is stationary or in motion, or whether or not the differential atmospheric delays (ionosphere and troposphere) are included as unknowns. An overview of

these models can be found in textbooks like [23.1–5] and in the Chaps. 21, 25, and 26 of this Handbook.

The theory that underpins ultraprecise GNSS carrier-phase ambiguity resolution is the theory of integer inference [23.6, 7]. This theory of integer estimation and validation is the topic of the present chapter. Although the theory was originally developed for Global Positioning System (GPS) [23.8–14], the theory has a much wider range of applicability. Next to the regional and global satellite navigation systems, it also applies to other carrier-phase-based interferometric techniques, such as Very Long Baseline Interferometry (VLBI) [23.15], Interferometric Synthetic Aperture Radar (InSAR) [23.16], or underwater acoustic carrier-phase positioning [23.17].

This chapter is organized as follows. In Sect. 23.1, the mixed-integer GNSS model is introduced. It forms the basis of all integer ambiguity resolution methods. An overview of the various ambiguity resolution steps is given, together with an evaluation of their contribution to the overall quality.

In Sect. 23.2, the ambiguity resolution methods of integer rounding (IR) and integer bootstrapping (IB) are presented, together with practical expressions for evaluating their ambiguity success rates. These methods are the simplest methods available, but their performance depends on the chosen ambiguity parametrization.

In Sect. 23.3 it is shown how the performance of rounding and bootstrapping can be improved by using certain ambiguity parametrizations. This includes a description of the decorrelating Z-transformation by which these improvements can be realized. Various examples that illustrate the concepts involved are also given.

The method of integer least-squares (ILSs) ambiguity resolution is described in Sect. 23.4. This method is optimal in the sense that it achieves the highest success

rate of all ambiguity resolution methods. The method is however also more complex as it requires an integer search over an ambiguity search space. It is shown how to make the method numerically efficient by combining the integer search with ambiguity decorrelation. Methods for computing or bounding the ILS success rate are also given.

The concept of partial ambiguity resolution is presented in Sect. 23.5. It is an alternative to full ambiguity resolution in case the resolution of all ambiguities cannot be done with a sufficiently high success rate.

As wrongly fixed integer ambiguities can result in unacceptably large positioning errors, it is important to have rigorous testing methods in place for accepting or rejecting the computed integer ambiguity solution. These methods and their theoretical foundation are presented in Sect. 23.6.

23.1 GNSS Ambiguity Resolution

23.1.1 The GNSS Model

To formulate the GNSS model for ambiguity resolution, we start with the observation equations for the pseudorange (code) and carrier-phase observables. If we denote the j -frequency pseudorange and carrier-phase for the r - s receiver-satellite combination at epoch t as $p_{r,j}^s(t)$ and $\phi_{r,j}^s(t)$ respectively, then their observation equations can be formulated as [23.1–5],

$$\begin{aligned} p_{r,j}^s(t) &= \rho_r^s(t) + T_r^s(t) + I_{r,j}^s(t) \\ &\quad + cd_{r,j}^s(t) + e_{r,j}^s(t), \\ \phi_{r,j}^s(t) &= \rho_r^s(t) + T_r^s(t) - I_{r,j}^s(t) \\ &\quad + c\delta t_{r,j}^s(t) + \lambda_j N_{r,j}^s + \epsilon_{r,j}^s(t), \end{aligned} \quad (23.1)$$

where ρ_r^s is the receiver-satellite range, $T_r^s(t)$ and $I_{r,j}^s$ are the tropospheric and ionospheric path delays, $d_{r,j}^s$ and $\delta t_{r,j}^s$ are the pseudorange and carrier-phase receiver-satellite clock biases, $N_{r,j}^s$ is the time-invariant integer carrier-phase ambiguity, c is the speed of light, λ_j is the j -frequency wavelength, and $e_{r,j}^s$, $\epsilon_{r,j}^s$ are the remaining error terms respectively.

The receiver-satellite range ρ_r^s in (23.1) is usually further linearized in the receiver- or satellite-position coordinates. As a result one obtains linear equations that can be used to form a system of linear equations for solving the unknown parameters of position, atmosphere, clock and ambiguities. Hence, if we assume the error terms $e_{r,j}^s$ and $\epsilon_{r,j}^s$ in (23.1) to be zero-mean random variables, the linear(ized) system of observation equations can be used to set up a linear model in

which some of the unknown parameters are reals and others are integer. Such a GNSS model is an example of a mixed-integer linear model.

We now define the general form of the mixed-integer GNSS model.

Definition 23.1 Mixed-integer GNSS model

Let (\mathbf{A}, \mathbf{B}) be a given $m \times (n+p)$ matrix of full rank and let \mathbf{Q}_{yy} be a given $m \times m$ positive definite matrix. Then

$$\mathbf{y} \sim \mathcal{N}(\mathbf{A}\mathbf{a} + \mathbf{B}\mathbf{b}, \mathbf{Q}_{yy}), \quad \mathbf{a} \in \mathbb{Z}^n, \quad \mathbf{b} \in \mathbb{R}^p \quad (23.2)$$

will be referred to as the mixed-integer GNSS model.

The notation \sim is used to describe *distributed as*. The m -vector \mathbf{y} contains the pseudorange and carrier-phase observables, the n -vector \mathbf{a} the integer ambiguities, and the real-valued p -vector \mathbf{b} the remaining unknown parameters, such as, for example, position coordinates, atmospheric delay parameters (troposphere, ionosphere) and clock parameters. As in most GNSS applications, the underlying probability distribution of the data is assumed to be a multivariate normal distribution.

23.1.2 Ambiguity Resolution Steps

The purpose of ambiguity resolution is to exploit the integer constraints, $\mathbf{a} \in \mathbb{Z}^n$ in (23.2), so as to get a better estimator of \mathbf{b} than otherwise would be the case. The mixed-integer GNSS model (23.2) can be solved in the following steps:

1. *Float Solution*: In the first step, the integer nature of the ambiguities is discarded and a standard least-squares (LS) parameter estimation is performed. As a result, one obtains the so-called float solution, together with its variance-covariance matrix,

$$\begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}} & \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{b}}} \\ \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{a}}} & \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{b}}} \end{bmatrix} \right). \quad (23.3)$$

Other forms than batch least-squares – such as recursive LS or Kalman filtering – may also be used to come up with a float solution. Such choices will depend on the application and on the structure of the GNSS model.

2. *Integer Solution*: The purpose of this second step is to take the integer constraints $\mathbf{a} \in \mathbb{Z}^n$ (23.2) into account. Hence, a mapping $\mathcal{I} : \mathbb{R}^n \mapsto \mathbb{Z}^n$ is introduced that maps the float ambiguities to corresponding integer values,

$$\check{\mathbf{a}} = \mathcal{I}(\hat{\mathbf{a}}). \quad (23.4)$$

Many such integer mappings \mathcal{I} exist. Popular choices are integer rounding (IR), integer bootstrapping (IB) and integer least-squares (ILS), see Sects. 23.2 and 23.4.

3. *Fixed Solution*: In the final step, once $\check{\mathbf{a}}$ is accepted, the ambiguity residual $\hat{\mathbf{a}} - \check{\mathbf{a}}$ is used to readjust the float estimator $\hat{\mathbf{b}}$ to obtain the so-called fixed esti-

mator

$$\check{\mathbf{b}} = \hat{\mathbf{b}} - \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{a}}} \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}^{-1} (\hat{\mathbf{a}} - \check{\mathbf{a}}). \quad (23.5)$$

The fixed solution has a quality that is commensurate with the high precision of the phase data, *provided* the probability of $\check{\mathbf{a}}$ being the correct integer is sufficiently high. Figure 23.1 illustrates the high gain in positioning precision that can be achieved with successful ambiguity resolution.

23.1.3 Ambiguity Resolution Quality

To determine the quality of the fixed solution $\check{\mathbf{b}}$ (23.5), we need to propagate the probabilistic properties of its constituents:

1. *Quality of float solution*: The float solution is defined as the minimizer of the unconstrained LS-problem,

$$(\hat{\mathbf{a}}, \hat{\mathbf{b}}) = \arg \min_{\mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{A}\mathbf{a} - \mathbf{B}\mathbf{b}\|_{\mathbf{Q}_{yy}}^2 \quad (23.6)$$

the solution of which follows from solving the normal equations

$$\begin{bmatrix} \mathbf{A}^\top \mathbf{Q}_{yy}^{-1} \mathbf{A} & \mathbf{A}^\top \mathbf{Q}_{yy}^{-1} \mathbf{B} \\ \mathbf{B}^\top \mathbf{Q}_{yy}^{-1} \mathbf{A} & \mathbf{B}^\top \mathbf{Q}_{yy}^{-1} \mathbf{B} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^\top \mathbf{Q}_{yy}^{-1} \mathbf{y} \\ \mathbf{B}^\top \mathbf{Q}_{yy}^{-1} \mathbf{y} \end{bmatrix}. \quad (23.7)$$

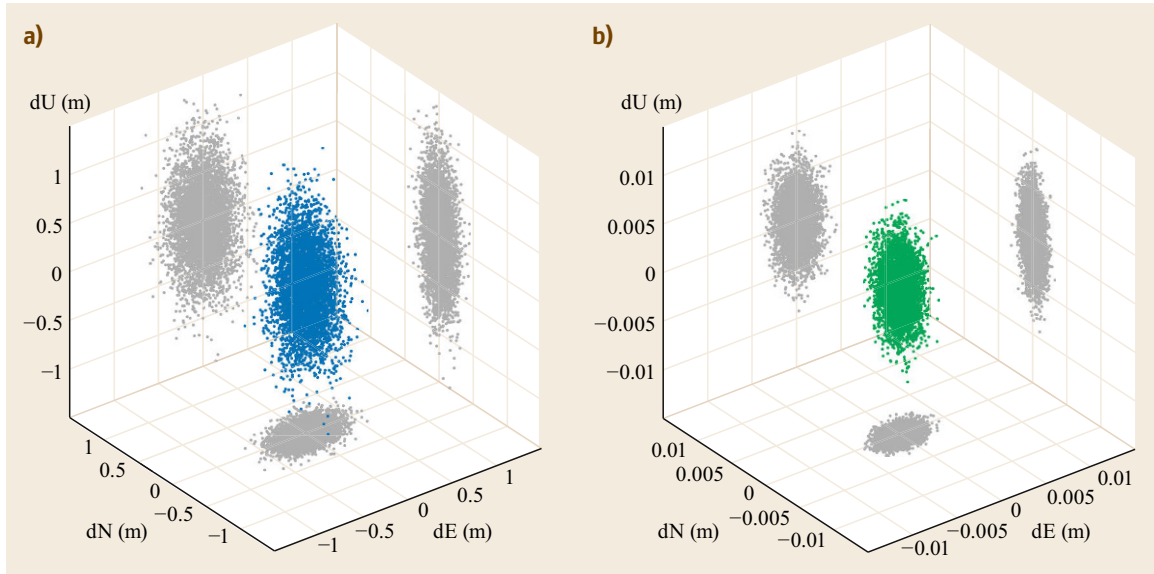


Fig. 23.1a,b Three-dimensional scatterplot of GPS position errors for short-baseline, dual-frequency instantaneous ambiguity *float* solutions ((a); $\hat{\mathbf{b}}$) and corresponding ambiguity *fixed* position solutions ((b); $\check{\mathbf{b}}$) (after [23.18]). Note the two orders of magnitude difference in scale between the two panels. dE, dN, and dU denote the components of the position errors in north, east and up direction

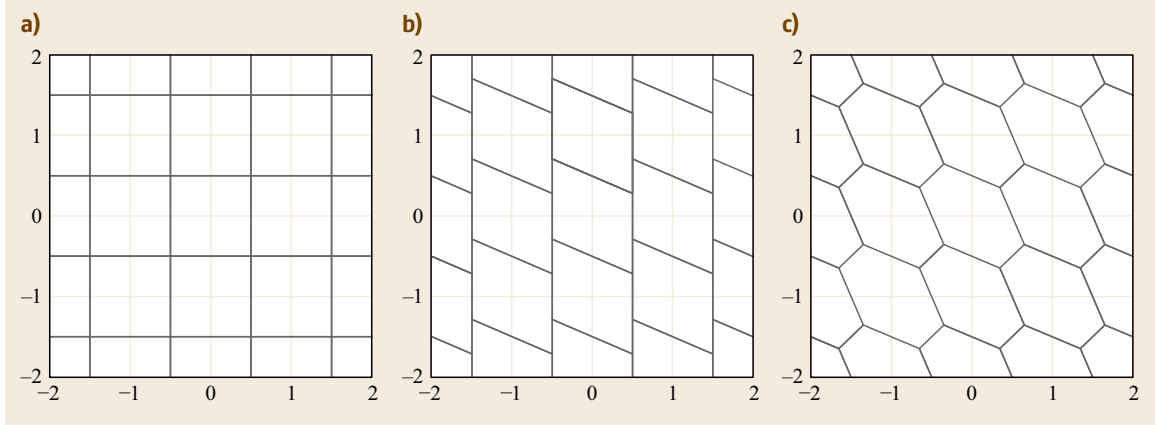


Fig. 23.2a–c Two-dimensional pull-in regions of integer rounding (a), integer bootstrapping (b), and integer least-squares (c)

This solution is given as

$$\begin{aligned}\hat{\mathbf{a}} &= (\bar{\mathbf{A}}^\top \mathbf{Q}_{yy}^{-1} \bar{\mathbf{A}})^{-1} \bar{\mathbf{A}}^\top \mathbf{Q}_{yy}^{-1} \mathbf{y} \\ \hat{\mathbf{b}} &= (\mathbf{B}^\top \mathbf{Q}_{yy}^{-1} \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{Q}_{yy}^{-1} (\mathbf{y} - \mathbf{A}\hat{\mathbf{a}}),\end{aligned}\quad (23.8)$$

where $\bar{\mathbf{A}} = \mathbf{P}_B^\perp \mathbf{A}$, with orthogonal projector

$$\mathbf{P}_B^\perp = \mathbf{I}_m - \mathbf{B}(\mathbf{B}^\top \mathbf{Q}_{yy}^{-1} \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{Q}_{yy}^{-1}.$$

With the distributional assumptions of (23.2), the distribution of the ambiguity float solution follows as the multivariate normal distribution $\hat{\mathbf{a}} \sim \mathcal{N}(\mathbf{a}, \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}})$, with variance matrix

$$\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}} = (\bar{\mathbf{A}}^\top \mathbf{Q}_{yy}^{-1} \bar{\mathbf{A}})^{-1}. \quad (23.9)$$

The probability density function (PDF) of $\hat{\mathbf{a}}$ is thus given as

$$\begin{aligned}f_{\hat{\mathbf{a}}}(\mathbf{x}|\mathbf{a}) &= \frac{1}{\sqrt{\det(2\pi\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}})}} \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{a}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2\right).\end{aligned}\quad (23.10)$$

Its shape is completely determined by the ambiguity variance matrix $\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}$, which in its turn is completely determined by the GNSS model's design matrix, (\mathbf{A}, \mathbf{B}) , and observation variance matrix \mathbf{Q}_{yy} . The PDF of $\hat{\mathbf{a}}$ is needed to determine the probability mass function (PMF) of $\check{\mathbf{a}}$ in step 2.

2. *Quality of integer solution:* Since the integer map of step 2, $\mathcal{I}: \mathbb{R}^n \mapsto \mathbb{Z}^n$, is a many-to-one map, different real-valued vectors will be mapped to one and the same integer vector. One can therefore assign a subset, say $\mathcal{P}_z \subset \mathbb{R}^n$, to each integer vector $\mathbf{z} \in \mathbb{Z}^n$,

$$\mathcal{P}_z = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{z} = \mathcal{I}(\mathbf{x})\}, \quad \mathbf{z} \in \mathbb{Z}^n. \quad (23.11)$$

This subset is referred to as the *pull-in region* of \mathbf{z} . It is the region in which all vectors are pulled to the same integer vector \mathbf{z} . The pull-in regions are translational invariant over the integers and cover the whole space \mathbb{R}^n without gaps and overlap [23.19]. Two-dimensional examples of pull-in regions are shown in Fig. 23.2. They are the pull-in regions of integer rounding, integer bootstrapping and integer least-squares.

The PMF of $\check{\mathbf{a}}$ follows from integrating the PDF of $\hat{\mathbf{a}}$ over the pull-in regions. Since $\check{\mathbf{a}} = \mathbf{z} \in \mathbb{Z}^n$ iff $\hat{\mathbf{a}} \in \mathcal{P}_z$, the PMF of $\check{\mathbf{a}}$ follows as

$$P(\check{\mathbf{a}} = \mathbf{z}) = P(\hat{\mathbf{a}} \in \mathcal{P}_z) = \int_{\mathcal{P}_z} f_{\hat{\mathbf{a}}}(\mathbf{x}|\mathbf{a}) d\mathbf{x}. \quad (23.12)$$

A two-dimensional example of an ambiguity PDF and corresponding PMF is given in Fig. 23.3a,b.

Of all the probabilities of the PMF, the probability of correct integer estimation, $P(\check{\mathbf{a}} = \mathbf{a})$, is of particular importance for ambiguity resolution. This probability is referred to as the ambiguity *success rate* and it is given by the integral

$$\begin{aligned}P_s = P(\check{\mathbf{a}} = \mathbf{a}) &= \int_{\mathcal{P}_a} f_{\hat{\mathbf{a}}}(\mathbf{x}|\mathbf{a}) d\mathbf{x} \\ &= \int_{\mathcal{P}_0} f_{\hat{\mathbf{a}}}(\mathbf{x}|\mathbf{0}) d\mathbf{x},\end{aligned}\quad (23.13)$$

where the last line follows from the translational property $f_{\hat{\mathbf{a}}}(\mathbf{x} + \mathbf{a}|\mathbf{a}) = f_{\hat{\mathbf{a}}}(\mathbf{x}|\mathbf{0})$ of the multivariate normal distribution.

Note that the success rate P_s depends on the pull-in region \mathcal{P}_0 and on the PDF $f_{\hat{\mathbf{a}}}(\mathbf{x}|\mathbf{0})$. Hence, the success rate is determined by the mapping $\mathcal{I}: \mathbb{R}^n \mapsto \mathbb{Z}^n$

Fig. 23.3 (a) Gaussian probability density function (PDF) $f_{\hat{a}}(\mathbf{x}|\mathbf{a})$ with 2-D (hexagon) ILS pull-in regions. (b) Corresponding probability mass function (PMF) $P(\check{\mathbf{a}}_{\text{ILS}} = \mathbf{z})$ of ILS estimator. (c) Scatterplot of horizontal position errors for float solution (gray dots) and corresponding fixed solution (green and red dots). In this case, 93% of the solutions were correctly fixed (green dots), and 7% were wrongly fixed (red dots) (after [23.18]) ►

and the ambiguity variance matrix $\mathbf{Q}_{\hat{a}\hat{a}}$, i. e., by the choice of integer estimator and the precision of the float ambiguities.

Due to the shape of the pull-in regions and the nondiagonality of the ambiguity variance matrix, the computation of the ambiguity success rate is nontrivial. The evaluation of the multivariate integral (23.13) can generally be done through Monte Carlo integration [23.20], see also Sect. 23.4.3. For some important integer estimators we also have easy-to-compute expressions and/or sharp (lower and upper) bounds of their success rates available (Sect. 23.2).

3. *Quality of fixed solution:* Once the integer solution is available, the fixed solution is computed as in (23.5). This fixed solution has the *multimodal* PDF [23.21]

$$f_{\hat{\mathbf{b}}}(\mathbf{x}) = \sum_{\mathbf{z} \in \mathbb{Z}^n} f_{\hat{\mathbf{b}}(\mathbf{z})}(\mathbf{x}) P(\check{\mathbf{a}} = \mathbf{z}) \quad (23.14)$$

in which $f_{\hat{\mathbf{b}}(\mathbf{z})}(\mathbf{x})$ denotes the PDF of the conditional LS-estimator

$$\hat{\mathbf{b}}(\mathbf{z}) = \hat{\mathbf{b}} - \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{a}}} \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}^{-1} (\hat{\mathbf{a}} - \mathbf{z}),$$

normally distributed with mean and variance matrix,

$$\begin{aligned} \mathbf{b}(\mathbf{z}) &= \mathbf{b} - \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{a}}} \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}^{-1} (\hat{\mathbf{a}} - \mathbf{z}), \\ \mathbf{Q}_{\hat{\mathbf{b}}(\mathbf{z})\hat{\mathbf{b}}(\mathbf{z})} &= \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{b}}} - \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{a}}} \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}^{-1} \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{b}}}. \end{aligned} \quad (23.15)$$

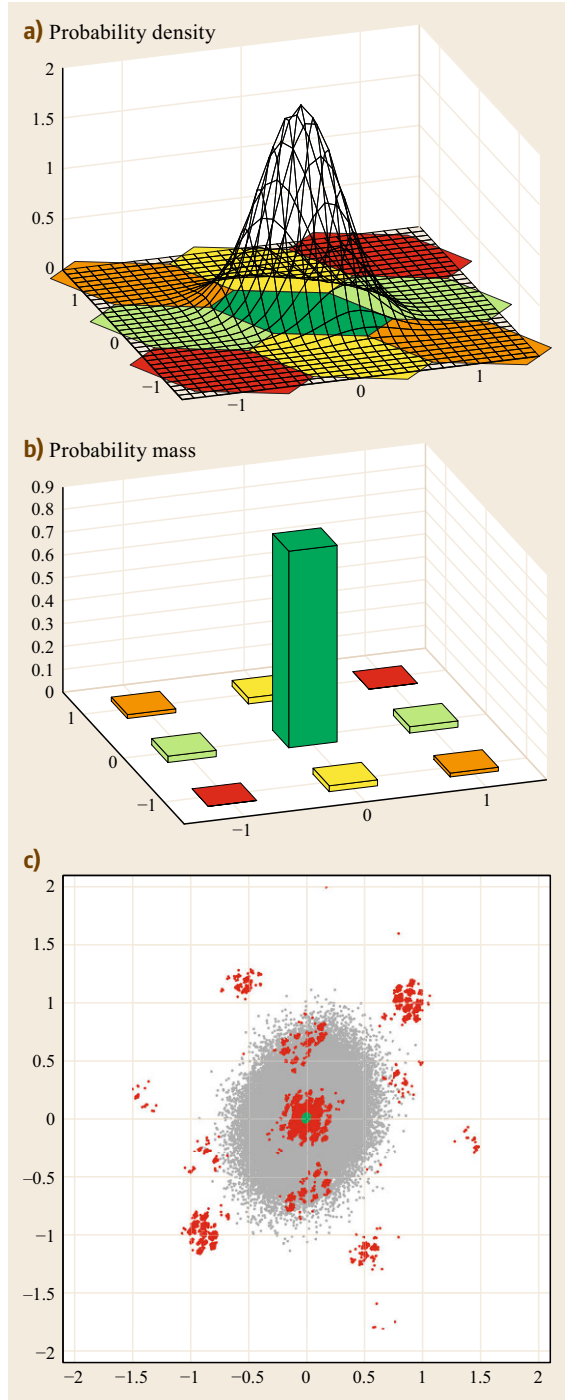
From (23.14) it follows that

$$f_{\hat{\mathbf{b}}}(\mathbf{x}) \approx f_{\hat{\mathbf{b}}(\mathbf{a})}(\mathbf{x}) \sim \mathcal{N}(\mathbf{b}, \mathbf{Q}_{\hat{\mathbf{b}}(\mathbf{z})\hat{\mathbf{b}}(\mathbf{z})}) \quad (23.16)$$

if

$$P_s = P(\check{\mathbf{a}} = \mathbf{a}) \approx 1. \quad (23.17)$$

Thus if the success rate is sufficiently close to one, the distribution of the fixed solution $\hat{\mathbf{b}}$ can be approximated by the unimodal normal distribution $\mathcal{N}(\mathbf{b}, \mathbf{Q}_{\hat{\mathbf{b}}(\mathbf{z})\hat{\mathbf{b}}(\mathbf{z})})$ of which the precision is better than that of the float solution $\hat{\mathbf{b}}$, $\mathbf{Q}_{\hat{\mathbf{b}}(\mathbf{z})\hat{\mathbf{b}}(\mathbf{z})} < \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{b}}}$.



The relevance of ambiguity resolution and the need to have sufficiently large success rates is illustrated in Fig. 23.3c. It shows scatterplots of float positions (gray scatter) and corresponding fixed positions (green/red scatter). The small size of the green scatter shows

the improvements that can be achieved over the float solution if the ambiguities are correctly fixed. The large red scatter indicates however that in this case the success rate is not large enough ($P_s = 93\%$) to

avoid some of the fixed positions being even poorer than the float positions. This underlines the importance of working with sufficiently high success rates only.

23.2 Rounding and Bootstrapping

23.2.1 Integer Rounding

The simplest integer estimator is *rounding to the nearest integer*. In the scalar case, its pull-in regions (intervals) are given as

$$\mathcal{R}_z = \left\{ x \in \mathbb{R} \mid |x - z| \leq \frac{1}{2} \right\}, \quad z \in \mathbb{Z}. \quad (23.18)$$

Any outcome of $\hat{a} \sim N(a \in \mathbb{Z}, \sigma_a^2)$, that satisfies $|\hat{a} - z| \leq 1/2$, will thus be pulled to the integer z . We denote the rounding estimator as \check{a}_R and the operation of integer rounding as $\lceil \cdot \rceil$. Thus $\check{a}_R = \lceil \hat{a} \rceil$ and $\check{a}_R = z$ if $\hat{a} \in \mathcal{R}_z$.

The PMF of $\check{a}_R = \lceil \hat{a} \rceil$ is given as

$$\begin{aligned} P(\check{a}_R = z) &= \left[\Phi \left(\frac{1 - 2(a - z)}{2\sigma_a} \right) + \Phi \left(\frac{1 + 2(a - z)}{2\sigma_a} \right) - 1 \right], \\ & \quad z \in \mathbb{Z}, \end{aligned} \quad (23.19)$$

where $\Phi(x)$ denotes the normal distribution function,

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}v^2\right) dv.$$

The PMF becomes more peaked when σ_a gets smaller. The success rate of scalar rounding follows from (23.19) by setting z equal to a ,

$$P(\check{a}_R = a) = 2\Phi\left(\frac{1}{2\sigma_a}\right) - 1. \quad (23.20)$$

The behavior of the success rate as function of the ambiguity standard deviation σ_a is shown in Fig. 23.4. It shows that a success rate better than 99%, requires $\sigma_a < 0.20$ cycle.

23.2.2 Vectorial Rounding

Scalar rounding is easily generalized to the vectorial case. It is defined as the componentwise rounding of

$\hat{\mathbf{a}} = (\hat{a}_1, \dots, \hat{a}_n)^\top$, $\check{\mathbf{a}}_R = (\lceil \hat{a}_1 \rceil, \lceil \hat{a}_2 \rceil, \dots, \lceil \hat{a}_n \rceil)^\top$. The pull-in regions of vectorial rounding are the multivariate versions of the scalar pull-in intervals,

$$\mathcal{R}_z = \left\{ \mathbf{x} \in \mathbb{R}^n \mid |\mathbf{c}_i^\top (\mathbf{x} - \mathbf{z})| \leq \frac{1}{2}, \quad i = 1, \dots, n \right\}, \quad (23.21)$$

with $\mathbf{z} \in \mathbb{Z}^n$ and where \mathbf{c}_i denotes the unit vector having a 1 as its i th entry and zeros otherwise. Thus the pull-in regions of rounding are unit-squares in two-dimensional (2-D), unit-cubes in three-dimensional (3-D), and so on (Fig. 23.2).

To determine the joint PMF of the components of $\check{\mathbf{a}}_R$, we have to integrate the PDF of $\hat{\mathbf{a}} \sim N(\mathbf{a}, \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}})$ over the pull-in regions \mathcal{R}_z . These n -fold integrals are difficult to evaluate unless the variance matrix $\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}$ is diagonal, in which case the components of $\check{\mathbf{a}}_R$ are independent and their joint PMF follows as the product of the univariate PMFs of the components. The corresponding success rate is then given by the n -fold product of the univariate success rates.

In case of GNSS, the variance matrix $\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}$ will be fully populated, meaning that one will have to resort to methods of Monte Carlo simulation for computing the joint PMF. For the success rate, one can alternatively make use of the following bounds.

Theorem 23.1 Rounding success-rate bounds [23.22]

Let the float ambiguity solution be distributed as $\hat{\mathbf{a}} \sim N(\mathbf{a}, \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}})$, $\mathbf{a} \in \mathbb{Z}^n$. Then the rounding success rate can be bounded from below and from above as

$$\text{LB} \leq P(\check{\mathbf{a}}_R = \mathbf{a}) \leq \text{UB}, \quad (23.22)$$

where

$$\begin{aligned} \text{LB} &= \prod_{i=1}^n \left[2\Phi\left(\frac{1}{2\sigma_{\hat{a}_i}}\right) - 1 \right], \\ \text{UB} &= \left[2\Phi\left(\frac{1}{2 \max_{i=1, \dots, n} \sigma_{\hat{a}_i}}\right) - 1 \right]. \end{aligned} \quad (23.23)$$

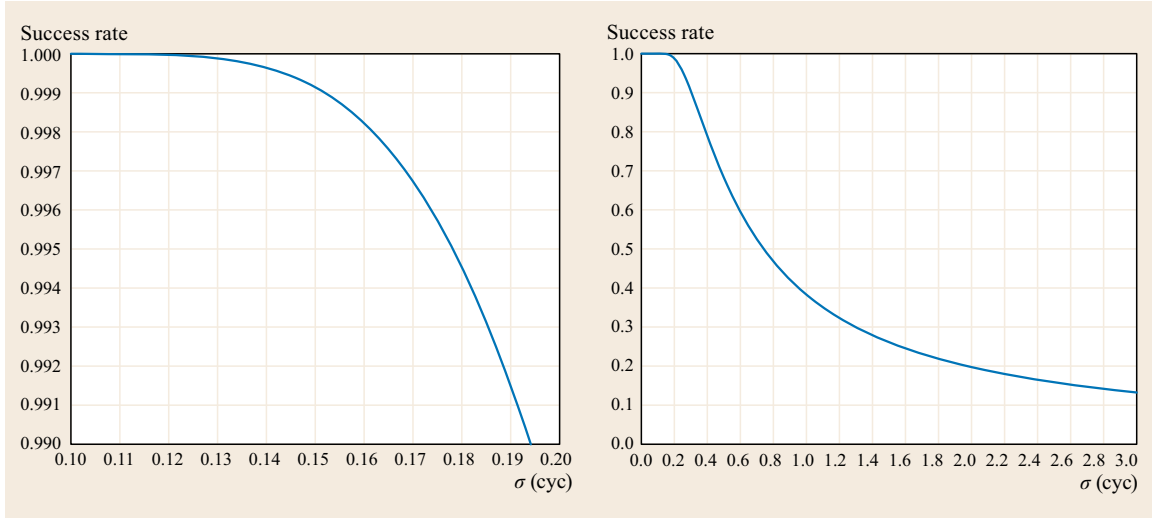


Fig. 23.4 Scalar rounding success rate versus ambiguity standard deviation σ in cycles

These easy-to-compute bounds are very useful for determining the expected success of GNSS ambiguity rounding. The upper bound is useful to quickly decide against such ambiguity resolution. It shows that ambiguity resolution based on vectorial rounding cannot be expected to be successful if already one of the scalar rounding success rates is too low.

The lower bound is useful to quickly decide in favor of vectorial rounding. If the lower bound is sufficiently close to 1, one can be confident that vectorial rounding will produce the correct integer ambiguity vector. Note that this requires each of the individual probabilities in the product of the lower bound to be sufficiently close to 1.

23.2.3 Integer Bootstrapping

Integer bootstrapping is a generalization of integer rounding; it combines integer rounding with sequential conditional least-squares estimation and as such takes some of the correlation between the components of the float solution into account. The method goes as follows. If $\hat{\mathbf{a}} = (\hat{a}_1, \dots, \hat{a}_n)^\top$, one starts with \hat{a}_1 and as before rounds its value to the nearest integer. Having obtained the integer of the first component, the real-valued estimates of all remaining components are then corrected by virtue of their correlation with \hat{a}_1 . Then the second, but now corrected, real-valued component is rounded to its nearest integer. Having obtained the integer value of this second component, the real-valued estimates of all remaining $n-2$ components are then again corrected by virtue of their correlation with the second component. This process is continued until all n components are taken care of. We have the following definition.

Definition 23.2 Integer bootstrapping

Let $\hat{\mathbf{a}} = (\hat{a}_1, \dots, \hat{a}_n)^\top \in \mathbb{R}^n$ be the float solution and let $\check{\mathbf{a}}_B = (\check{a}_{B,1}, \dots, \check{a}_{B,n})^\top \in \mathbb{Z}^n$ denote the corresponding integer bootstrapped solution. Then

$$\begin{aligned} \check{a}_{B,1} &= \lceil \hat{a}_1 \rceil, \\ \check{a}_{B,2} &= \lceil \hat{a}_{2|1} \rceil = \lceil \hat{a}_2 - \sigma_{21}\sigma_1^{-2}(\hat{a}_1 - \check{a}_{B,1}) \rceil, \\ &\vdots \\ \check{a}_{B,n} &= \lceil \hat{a}_{n|N} \rceil = \lceil \hat{a}_n - \sum_{j=1}^{n-1} \sigma_{n,j|J} \sigma_{j|J}^{-2}(\hat{a}_{j|J} - \check{a}_{B,j}) \rceil, \end{aligned} \quad (23.24)$$

where $\hat{a}_{i|I}$ is the least-squares estimator of a_i conditioned on the values of the previous $I = \{1, \dots, (i-1)\}$ sequentially rounded components, $\sigma_{i,j|J}$ is the covariance between \hat{a}_i and $\hat{a}_{j|J}$, and $\sigma_{j|J}^2$ is the variance of $\hat{a}_{j|J}$. For $i = 1$, $\hat{a}_{i|I} = \hat{a}_1$.

As the definition shows, the bootstrapped estimator can be seen as a generalization of integer rounding. The bootstrapped estimator reduces to integer rounding in the case where correlations are absent, i.e., in the case where the variance matrix $Q_{\hat{\mathbf{a}}}$ is diagonal.

In vector-matrix form, the bootstrapped estimator (23.24) can shown to be given as [23.23],

$$\check{\mathbf{a}}_B = \lceil \hat{\mathbf{a}} + (\mathbf{L}^{-1} - \mathbf{I}_n)(\hat{\mathbf{a}} - \check{\mathbf{a}}_B) \rceil, \quad (23.25)$$

with \mathbf{L} the unit lower triangular matrix of the triangular decomposition $\mathbf{Q}_{\hat{a}\hat{a}} = \mathbf{L}\mathbf{D}\mathbf{L}^\top$. As the diagonal matrix

$$\mathbf{D} = \text{diag}(\sigma_{a_1}^2, \dots, \sigma_{a_{|N|}}^2)$$

is not used in the construction of the bootstrapped estimator, bootstrapping takes only part of the information of the variance matrix into account. Although the diagonal matrix \mathbf{D} is not used in (23.25), it is needed to determine the bootstrapped success rate.

23.2.4 Bootstrapped Success Rate

To determine the bootstrapped PMF, we first need the bootstrapped pull-in regions. They are given as

$$\mathcal{B}_z = \left\{ \mathbf{x} \in \mathbb{R}^n \mid |\mathbf{c}_i^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{z})| \leq \frac{1}{2}, i = 1, \dots, n \right\}, \quad (23.26)$$

with $\mathbf{z} \in \mathbb{Z}^n$ and where \mathbf{c}_i denotes the unit vector having a 1 as its i th entry and zeros otherwise. They are parallelograms in 2-D (Fig. 23.2).

The bootstrapped PMF follows from integrating the multivariate normal distribution over the bootstrapped pull-in regions. In contrast to the multivariate integral for integer rounding, the multivariate integral for bootstrapping can be simplified considerably. As shown by the following theorem, the bootstrapped PMF can be expressed as a product of univariate integrals.

Theorem 23.2 Bootstrapped PMF [23.22]

Let $\hat{\mathbf{a}} \sim \mathcal{N}(\mathbf{a} \in \mathbb{Z}^n, \mathbf{Q}_{\hat{a}\hat{a}})$ and let $\check{\mathbf{a}}_B$ be the bootstrapped estimator of \mathbf{a} . Then

$$P(\check{\mathbf{a}}_B = \mathbf{z}) = \prod_{i=1}^n \left[\Phi \left(\frac{1 - 2\mathbf{I}_i^\top (\mathbf{a} - \mathbf{z})}{2\sigma_{\hat{a}_{i|I}}} \right) + \Phi \left(\frac{1 + 2\mathbf{I}_i^\top (\mathbf{a} - \mathbf{z})}{2\sigma_{\hat{a}_{i|I}}} \right) - 1 \right], \quad (23.27)$$

with $\mathbf{z} \in \mathbb{Z}^n$ and where \mathbf{I}_i is the i th column vector of the unit upper triangular matrix $(\mathbf{L}^{-1})^\top$.

As a direct consequence of the above theorem, we have an exact and easy-to-compute expression for the bootstrapped success rate.

Corollary 23.1 Bootstrapped success rate

Let $\hat{\mathbf{a}} \sim \mathcal{N}(\mathbf{a} \in \mathbb{Z}^n, \mathbf{Q}_{\hat{a}\hat{a}})$. Then the bootstrapped success rate is given as

$$P(\check{\mathbf{a}}_B = \mathbf{a}) = \prod_{i=1}^n \left[2\Phi \left(\frac{1}{2\sigma_{\hat{a}_{i|I}}} \right) - 1 \right]. \quad (23.28)$$

This is an important result as it provides a simple way for evaluating the bootstrapped success rate.

When comparing the performance of bootstrapping with rounding, it can be shown that the success rate of bootstrapping will never be smaller than that of rounding [23.22],

$$P(\check{\mathbf{a}}_B = \mathbf{a}) \geq P(\check{\mathbf{a}}_R = \mathbf{a}). \quad (23.29)$$

Thus bootstrapping is a better integer estimator than rounding.

Despite the fact that we have the above exact and easy-to-compute formula for the bootstrapped success rate, an easy-to-compute upper bound of it would still be useful if it would be \mathbf{Z} -invariant. Such an upper bound can be constructed when use is made of the \mathbf{Z} -invariant ADOP (ambiguity dilution of precision).

Theorem 23.3 Bootstrapped success-rate invariant upper bound [23.24]

Let $\hat{\mathbf{a}} \sim \mathcal{N}(\mathbf{a}, \mathbf{Q}_{\hat{a}\hat{a}})$, $\mathbf{a} \in \mathbb{Z}^n$, $\hat{\mathbf{z}} = \mathbf{Z}\hat{\mathbf{a}}$ and $\text{ADOP} = \det(\mathbf{Q}_{\hat{a}\hat{a}})^{\frac{1}{2n}}$. Then

$$P(\check{\mathbf{z}}_B = \mathbf{z}) \leq \left[2\Phi \left(\frac{1}{2\text{ADOP}} \right) - 1 \right]^n \quad (23.30)$$

for any admissible \mathbf{Z} -transformation.

Thus if the upper bound is too small, we can immediately conclude, for any ambiguity parametrization, that bootstrapping nor rounding will be successful.

23.3 Linear Combinations

23.3.1 Z-transformations

Although the integer estimators $\check{\mathbf{a}}_R$ and $\check{\mathbf{a}}_B$ are easy to compute, they both suffer from a lack of invariance against integer reparametrizations or so-called Z-transformations.

Definition 23.3 Z-transformations [23.25]

An $n \times n$ matrix \mathbf{Z} is called a Z-transformation iff $\mathbf{Z}, \mathbf{Z}^{-1} \in \mathbb{Z}^{n \times n}$, i. e., if the entries of the matrix and its inverse are all integer.

Z-transformations leave the integer nature of integer vectors invariant. It can be shown that the two conditions, $\mathbf{Z}, \mathbf{Z}^{-1} \in \mathbb{Z}^{n \times n}$, are equivalent to the two conditions $\mathbf{Z} \in \mathbb{Z}^{n \times n}$ and $\det(\mathbf{Z}) = \pm 1$. Hence, the class of Z-transformations can also be defined as

$$\mathbf{Z} = \{\mathbf{Z} \in \mathbb{Z}^{n \times n} \mid |\mathbf{Z}| = \pm 1\}. \quad (23.31)$$

Thus, Z-transformations are volume-preserving transformations. This implies that the determinant of the ambiguity variance matrix is invariant for Z-transformations: $|\mathbf{Q}_{\hat{\mathbf{z}}\hat{\mathbf{z}}}| = |\mathbf{Z}\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}\mathbf{Z}^T| = |\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}|$.

By saying that an estimator lacks Z-invariance, we mean that if the float solution is Z-transformed, the integer solution does not transform accordingly. That is, rounding/bootstrapping and transforming do generally not commute,

$$\check{\mathbf{z}}_R \neq \mathbf{Z}\check{\mathbf{a}}_R \text{ and } \check{\mathbf{z}}_B \neq \mathbf{Z}\check{\mathbf{a}}_B \text{ if } \hat{\mathbf{z}} = \mathbf{Z}\hat{\mathbf{a}}. \quad (23.32)$$

This is illustrated in Fig. 23.5 for integer rounding and in Fig. 23.6 for integer bootstrapping. Also the success

rates of rounding and bootstrapping lack Z-invariance,

$$\begin{aligned} P(\check{\mathbf{z}}_R = \mathbf{z}) &\neq P(\check{\mathbf{a}}_R = \mathbf{a}), \\ P(\check{\mathbf{z}}_B = \mathbf{z}) &\neq P(\check{\mathbf{a}}_B = \mathbf{a}). \end{aligned} \quad (23.33)$$

This is also very clear from Figs. 23.5 and 23.6. Since the scatterplot of $\hat{\mathbf{a}}$ is much more elongated than that of $\hat{\mathbf{z}} = \mathbf{Z}\hat{\mathbf{a}}$, the rounding pull-in region is a much poorer fit of the original scatterplot than of the transformed scatterplot. This is also true for the bootstrapped pull-in regions, even though the shape of the bootstrapped pull-in region changes with the Z-transformation. Note that the two figures also illustrate the workings of inequality (23.29), i. e., that bootstrapping outperforms rounding. The bootstrapped pull-in regions have a better fit of the scatterplot, original as well as transformed, than the pull-in region of rounding.

The question is now whether the above-identified lack of invariance means that rounding and bootstrapping are unfit for GNSS integer ambiguity resolution? The answer is no, by no means. Integer rounding and bootstrapping are valid ambiguity estimators, and they are attractive, because of their computational simplicity. Whether or not they can be successfully applied in any concrete situation, depends solely on the value of their success rates for that particular situation.

23.3.2 (Extra) Widelaning

Since the performance of rounding and bootstrapping depends on the chosen ambiguity parameterization, it would be helpful to know how to improve their performance by choosing suitable Z-transformations. The simplest such Z-transformations are the so-called widelaning transformations. Examples of widelaning trans-

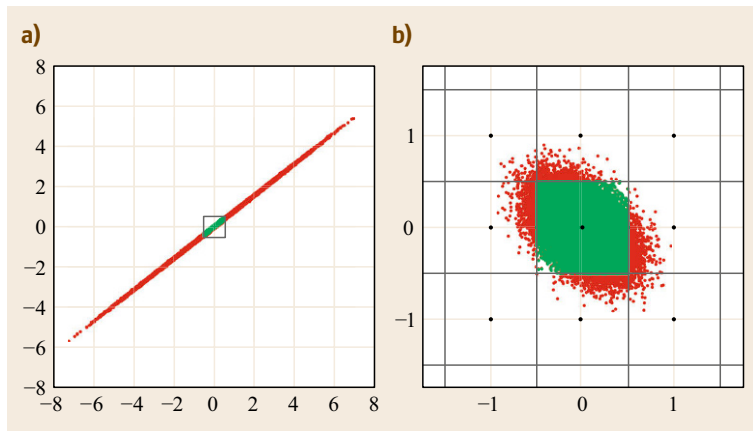


Fig. 23.5a,b 2-D IR pull-in regions and 50 000 simulated zero-mean float solutions. (a) Original ambiguities $\hat{\mathbf{a}}$ [cycles]; (b) Z-transformed ambiguities $\hat{\mathbf{z}} = \mathbf{Z}\hat{\mathbf{a}}$ [cycles]. Red dots will be pulled to wrong integer solutions, while green dots will be pulled to the correct integer solution (after [23.18])

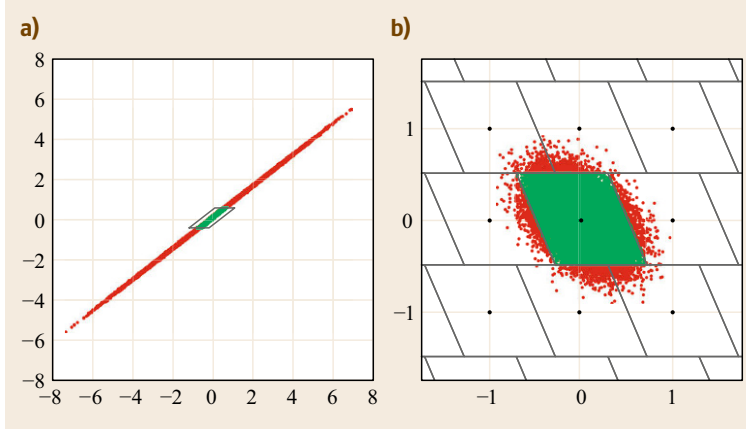


Fig. 23.6 2-D IB pull-in regions (original and transformed) and 50 000 simulated zero-mean float solutions. **(a)** Original ambiguities \hat{a} [cycles]; **(b)** Z-transformed ambiguities $\hat{z} = \mathbf{Z}\hat{a}$ [cycles]. Red dots will be pulled to wrong integer solutions, while green dots will be pulled to the correct integer solution (after [23.18])

formations are

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad (23.34)$$

for the dual-frequency case, and

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & -1 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \quad (23.35)$$

for the triple-frequency case. These transformations are referred to as *widelaning*, since they can be interpreted to form carrier-phase observables with long wavelengths. To see this, consider the carrier-phase transformation

$$\bar{\phi}_i = \frac{\sum_{j=1}^f Z_{ij} \lambda_j^{-1} \phi_j}{\sum_{j=1}^f Z_{ij} \lambda_j^{-1}}, \quad i = 1, \dots, f, \quad (23.36)$$

in which ϕ_j denotes the double-differenced (DD) carrier-phase observable on frequency $j = 1, \dots, f$, λ_j its wavelength and Z_{ij} the ij th-entry of the Z-transformation matrix. With this transformation, the system of f DD carrier-phase observation equations

$$\phi_i = \rho - \mu_i I + \lambda_i a_i, \quad i = 1, \dots, f \quad (23.37)$$

transforms to a system with similar structure, namely

$$\bar{\phi}_i = \rho - \bar{\mu}_i I + \bar{\lambda}_i z_i, \quad i = 1, \dots, f, \quad (23.38)$$

with ρ the DD nondispersive range plus tropospheric delay, I the DD ionospheric delay on the first frequency, μ_i and $\bar{\mu}_i$ the original and transformed ionospheric coefficient, λ_i and $\bar{\lambda}_i$ the original and transformed wavelength, and a_i and z_i the original and transformed ambiguity.

The relation between the original and transformed wavelengths is given as

$$\bar{\lambda}_i = \left(\sum_{j=1}^f Z_{ij} \lambda_j^{-1} \right)^{-1}, \quad i = 1, \dots, f. \quad (23.39)$$

If we now substitute the entries of the above widelaning transformation (23.35), together with the wavelengths of GPS (or Galileo or BeiDou), we obtain the values of the transformed wavelengths ($\bar{\lambda}_1 = \lambda_{\text{ew}}$, $\bar{\lambda}_2 = \lambda_{\text{w}}$, $\bar{\lambda}_3 = \lambda_3$) as given in Table 23.1, which indeed are larger than the original wavelengths.

The rationale for aiming at longer wavelengths is that a larger ambiguity coefficient $\bar{\lambda}_i$ improves the precision with which the ambiguity z_i can be estimated. However, this reasoning is only valid of course if all other circumstances remain unchanged under the transformation. This is not really the case with the above carrier-phase transformation (23.36), since the variance matrix of ϕ_i , $i = 1, \dots, f$, will generally differ from that of the transformed $\bar{\phi}_i$, $i = 1, \dots, f$. Nevertheless, the above simple widelaning transformations, (23.34) and (23.35), are still useful as they can often be seen as an easy first step in improving the precision of the float ambiguities.

23.3.3 Decorrelating Transformation

In general the widelaning approach is quite limited in finding suitable Z-transformations. We now describe a general method, due to [23.14], for finding such transformations. The method can be applied to any possible integer GNSS model and it has generally a significantly improved performance over widelaning [23.26–28].

Since it is the ambiguity variance matrix that completely determines the ambiguity success rate (23.13), the method takes the ambiguity variance matrix $\mathbf{Q}_{\hat{a}\hat{a}}$

Table 23.1 GPS, Galileo and BeiDou original, widelane (w) and extra widelane (ew) wavelengths (cm)

Wavelength	GPS		Galileo		BeiDou	
a_1	L_1	19.0	E_1	19.0	B_1	19.2
a_2	L_2	24.4	E_6	23.4	B_3	23.6
$z_3 = a_3$	L_5	25.5	E_{5a}	25.5	B_2	24.8
$z_2 = a_1 - a_2$	L_w	86.2	E_w	101.1	B_w	102.4
$z_1 = a_2 - a_3$	L_{ew}	587.0	E_{ew}	292.8	B_{ew}	488.9

as its point of departure. The aim is to find a Z-transformation that decorrelates the ambiguities as much as possible, i. e., that makes the transformed ambiguity variance matrix $\mathbf{Q}_{\hat{z}\hat{z}} = \mathbf{Z}\mathbf{Q}_{\hat{a}\hat{a}}\mathbf{Z}^\top$ as diagonal as possible. The rationale of this approach is that an ambiguity parametrization with diagonal variance matrix is optimal in the sense that then no further success-rate improvements of rounding and bootstrapping are possible through reparametrization.

The degree of decorrelation of a variance matrix is measured by its decorrelation number. Let

$$\mathbf{R}_{\hat{a}\hat{a}} = [\text{diag}(\mathbf{Q}_{\hat{a}\hat{a}})]^{-1/2} \mathbf{Q}_{\hat{a}\hat{a}} [\text{diag}(\mathbf{Q}_{\hat{a}\hat{a}})]^{-1/2}$$

be the correlation matrix of $\mathbf{Q}_{\hat{a}\hat{a}}$. Then the *decorrelation number* is defined as [23.26]

$$r_{\hat{a}} = \sqrt{|\mathbf{R}_{\hat{a}\hat{a}}|} \quad (0 \leq r_{\hat{a}} \leq 1). \quad (23.40)$$

In two dimensions it reduces to

$$r_{\hat{a}} = \sqrt{(1 - \rho_{\hat{a}}^2)},$$

with $\rho_{\hat{a}}$ being the ambiguity correlation coefficient. Hence, a two-dimensional ambiguity variance matrix is diagonal if and only if $r_{\hat{a}} = 1$. It can be shown that this also holds true for the higher-dimensional case. Since $|\mathbf{R}_{\hat{a}\hat{a}}| = |\mathbf{Q}_{\hat{a}\hat{a}}| / (\prod_{i=1}^n \sigma_{\hat{a}_i}^2)$ and $|\mathbf{Q}_{\hat{z}\hat{z}}| = |\mathbf{Q}_{\hat{a}\hat{a}}|$, we have

$$r_{\hat{z}} \geq r_{\hat{a}} \Leftrightarrow \sigma_{\hat{z}_1}^2 \dots \sigma_{\hat{z}_n}^2 \leq \sigma_{\hat{a}_1}^2 \dots \sigma_{\hat{a}_n}^2. \quad (23.41)$$

Hence, the ambiguity decorrelation number increases if the product of ambiguity variances decreases. We now show how to construct such decorrelating Z-transformation for the two-dimensional case. For the higher-dimensional case, see for example [23.27, 28].

We minimize the product $\sigma_{\hat{a}_1}^2 \sigma_{\hat{a}_2}^2$ in an alternating fashion, i. e., we start by keeping the first variance unchanged and reduce the second variance. Then we keep the second, now reduced, variance unchanged and reduce the first variance. This process is continued until no further reduction in the product of variances is possible anymore.

In the sequence of alternating reductions, the following type of transformations are applied

$$\mathbf{\Pi}_2 \mathbf{G}_\alpha = \begin{bmatrix} \alpha & 1 \\ 1 & 0 \end{bmatrix}, \quad (23.42)$$

where

$$\mathbf{G}_\alpha = \begin{bmatrix} 1 & 0 \\ \alpha & 1 \end{bmatrix}, \quad \mathbf{\Pi}_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \quad (23.43)$$

With \mathbf{G}_α , the variance of the second ambiguity is reduced, while with $\mathbf{\Pi}_2$, the order of the two ambiguities is interchanged. Once the order is interchanged, a transformation like \mathbf{G}_α can again be applied to further reduce the product of variances.

The value of α is determined in each step of the sequence as follows. With \mathbf{G}_α , the variance of the second ambiguity becomes

$$\begin{aligned} [\mathbf{G}_\alpha \mathbf{Q}_{\hat{a}\hat{a}} \mathbf{G}_\alpha^\top]_{22} &= \alpha^2 \sigma_{\hat{a}_1}^2 + 2\alpha \sigma_{\hat{a}_2 \hat{a}_1} + \sigma_{\hat{a}_2}^2 \\ &= \sigma_{\hat{a}_2}^2 - \sigma_{\hat{a}_1}^2 [(\sigma_{\hat{a}_2 \hat{a}_1} \sigma_{\hat{a}_1}^{-2})^2 - (\alpha + \sigma_{\hat{a}_2 \hat{a}_1} \sigma_{\hat{a}_1}^{-2})^2]. \end{aligned} \quad (23.44)$$

This shows that the variance of the transformed ambiguity is minimal for $\alpha = -\sigma_{\hat{a}_2 \hat{a}_1} \sigma_{\hat{a}_1}^{-2}$. As this is not an integer in general, it would not produce an admissible transformation when substituted into \mathbf{G}_α of (23.43). Therefore, instead of using the real-valued minimizer $-\sigma_{\hat{a}_2 \hat{a}_1} \sigma_{\hat{a}_1}^{-2}$ for α in \mathbf{G}_α , its nearest integer is used as approximation, $\alpha = -\lceil \sigma_{\hat{a}_2 \hat{a}_1} \sigma_{\hat{a}_1}^{-2} \rceil$. This still gives a reduction in the variance of the second ambiguity, since

$$(\sigma_{\hat{a}_2 \hat{a}_1} \sigma_{\hat{a}_1}^{-2})^2 > (\alpha + \sigma_{\hat{a}_2 \hat{a}_1} \sigma_{\hat{a}_1}^{-2})^2$$

if

$$|\sigma_{\hat{a}_2 \hat{a}_1} \sigma_{\hat{a}_1}^{-2}| > \frac{1}{2},$$

i. e., if

$$\lceil \sigma_{\hat{a}_2 \hat{a}_1} \sigma_{\hat{a}_1}^{-2} \rceil \neq 0.$$

The construction of the decorrelation transformation is summarized in the following definition [23.26–28].

Definition 23.4 Decorrelating Z-transformation

Let $\mathbf{Q}^{(1)} = \mathbf{Q}_{\hat{a}\hat{a}}$ and $\mathbf{Q}^{(i+1)} = \mathbf{Z}_i \mathbf{Q}^{(i)} \mathbf{Z}_i^\top$, $i = 1, \dots, k+2$. Then the two-dimensional decorrelating Z-transformation is given as the product

$$\mathbf{Z} = \mathbf{Z}_k \mathbf{Z}_{k-1} \dots \mathbf{Z}_1, \quad (23.45)$$

Table 23.2 2-D example of rounding and bootstrapping on original and Z-transformed ambiguities

$\mathbf{Z} = \begin{bmatrix} 4 & -3 \\ -1 & 1 \end{bmatrix}$	$\hat{\mathbf{a}} = \begin{bmatrix} 2.23 \\ 2.51 \end{bmatrix}$, $\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}} = \begin{bmatrix} 0.1680 & 0.2152 \\ 0.2152 & 0.2767 \end{bmatrix}$ Original ambiguities, $\rho_{\hat{\mathbf{a}}} = 0.96$	$\hat{\mathbf{z}} = \begin{bmatrix} 1.39 \\ 0.28 \end{bmatrix}$, $\mathbf{Q}_{\hat{\mathbf{z}}\hat{\mathbf{z}}} = \begin{bmatrix} 0.0135 & 0.0043 \\ 0.0043 & 0.0143 \end{bmatrix}$ Transformed ambiguities, $\rho_{\hat{\mathbf{z}}} = 0.31$
Rounding	$\check{\mathbf{a}}_{\text{R}} = [2, 3]^{\top}$	$\check{\mathbf{z}}_{\text{R}} = [1, 0]^{\top}$
Bootstrapping	$\check{\mathbf{a}}_{\text{B}}^{(1)} = [2, 2]^{\top}$, $\check{\mathbf{a}}_{\text{B}}^{(2)} = [3, 3]^{\top}$	$\check{\mathbf{z}}_{\text{B}}^{(1)} = [1, 0]^{\top}$, $\check{\mathbf{z}}_{\text{B}}^{(2)} = [1, 0]^{\top}$

where

$$\mathbf{Z}_i = \begin{bmatrix} \alpha_i & 1 \\ 1 & 0 \end{bmatrix}, \mathbf{Q}^{(i)} = \begin{bmatrix} \sigma_1^2(i) & \sigma_{12}(i) \\ \sigma_{21}(i) & \sigma_2^2(i) \end{bmatrix}$$

and

$$\alpha_i = -[\sigma_{21}(i)\sigma_1^{-2}(i)],$$

with $\alpha_{k+1} = \alpha_{k+2} = 0$.

After the above decorrelating transformation is applied, the correlation coefficient of the transformed ambiguities will never be larger than 0.5 in absolute value. This can be seen as follows. If $\alpha_{k+1} = \alpha_{k+2} = 0$, then $\sigma_{21}(k+2) = \sigma_{21}(k+1)$ and $\sigma_1^2(k+2) = \sigma_2^2(k+1)$, and therefore

$$\rho_z^2 = \frac{\sigma_{21}(k+1)^2}{\sigma_1^2(k+1)\sigma_2^2(k+1)} \leq \frac{1}{4}. \quad (23.46)$$

Geometrically, the above sequence of transformations in the product of \mathbf{Z} (23.45) can be given the following useful interpretation. Consider the confidence ellipse of $\hat{\mathbf{a}}$. Its shape and orientation is determined by $\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}$. The part \mathbf{G}_{α_1} of \mathbf{Z}_1 then pushes the two horizontal tangents of the ellipse inwards, while at the same time keeping fixed the area of the ellipse and the location of the two vertical tangents. Then $\mathbf{G}_{\alpha_2}\mathbf{\Pi}_2$ of the product $\mathbf{Z}_2\mathbf{Z}_1$ pushes the two vertical tangents of the ellipse inwards, while at the same time keeping fixed the area of the ellipse and the location of the two horizontal tangents. This process is continued until no further reduction is possible. Since the area of ellipse is kept constant at all times ($|\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}| = |\mathbf{Q}_{\hat{\mathbf{z}}\hat{\mathbf{z}}}|$), whereas the area of the enclosing rectangular box is reduced in each step, it follows that not only the diagonality of the ambiguity variance matrix is reduced, but also that the shape of the ellipse is forced to become more circular.

For further computational details on how such Z-transformations can be constructed, we refer to [23.14, 27–29] and the references cited therein. Also see [23.30–34].

23.3.4 Numerical Example

The following two-dimensional numerical example compares rounding with bootstrapping and illustrates

their dependence on the chosen ambiguity parametrization. The float solution has been computed from a dual-frequency, ionosphere-fixed geometry-free model for two receivers, two satellites, and two epochs, in which an undifferenced phase standard deviation of 3 mm and an undifferenced code standard deviation of 10 cm is assumed.

Hence, the computations are based on the double-differenced (DD) phase- and code-observation equations

$$\begin{aligned} \phi_i(t) &= \rho(t) + \lambda_i a_i + e_{\phi_i}(t), \\ p_i(t) &= \rho(t) + e_{p_i}(t), \end{aligned} \quad (23.47)$$

with $i = 1, 2$ and $t = t_1, t_2$.

The original and transformed float solution, $\hat{\mathbf{a}}$ and $\hat{\mathbf{z}} = \mathbf{Z}\hat{\mathbf{a}}$, and their variance matrices, $\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}$ and $\mathbf{Q}_{\hat{\mathbf{z}}\hat{\mathbf{z}}}$, are given in Table 23.2, together with the decorrelating transformation matrix \mathbf{Z} . It is constructed as

$$\mathbf{Z} = \begin{bmatrix} -3 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 4 & -3 \\ -1 & 1 \end{bmatrix}. \quad (23.48)$$

This transformation decorrelates ($\rho_{\hat{\mathbf{a}}} = 0.96$ versus $\rho_{\hat{\mathbf{z}}} = 0.31$) and significantly improves the precision of the ambiguities (Table 23.2). Also note that the first step in the construction of \mathbf{Z} consists of widening.

Table 23.2 also contains six integer solutions, two based on rounding and four based on bootstrapping. Rounding of $\hat{\mathbf{a}}$ gives

$$\check{\mathbf{a}}_{\text{R}} = \begin{bmatrix} \lceil 2.23 \rceil \\ \lceil 2.51 \rceil \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix},$$

while bootstrapping of $\hat{\mathbf{a}}$ gives

$$\check{\mathbf{a}}_{\text{B}}^{(1)} = \begin{bmatrix} \lceil 2.23 \rceil \\ \lceil 2.51 - \frac{0.2152}{0.1680}(2.23 - 2) \rceil \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix},$$

when starting from the first ambiguity, and

$$\check{\mathbf{a}}_{\text{B}}^{(2)} = \begin{bmatrix} \lceil 2.23 - \frac{0.2152}{0.2767}(2.51 - 3) \rceil \\ \lceil 2.51 \rceil \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix},$$

when starting from the second ambiguity. These solutions together with their counterparts in the transformed domain can be found in Table 23.2.

Note that all three solutions in the original domain, $\check{\mathbf{a}}_{\text{R}}$, $\check{\mathbf{a}}_{\text{B}}^{(1)}$ and $\check{\mathbf{a}}_{\text{B}}^{(2)}$, are different, while their counterparts in the transformed domain are the same and all equal to $[1, 0]^T$. Also note that when the solution in the transformed domain is back-transformed to the original domain, again a different solution is obtained, namely,

$$\check{\mathbf{a}}'_{\text{R}} = \mathbf{Z}^{-1}\check{\mathbf{z}}_{\text{R}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \quad (23.49)$$

In Table 23.3, the success rates of the different solutions are given. Note the big differences between the success rates of the transformed ambiguities and original ambiguities. The success rates of the transformed ambiguities are all very close to 1. This is due to the high precision of the transformed float solution $\hat{\mathbf{z}}$ (Table 23.2). Also note that the success rates of $\check{\mathbf{a}}_{\text{B}}^{(1)}$ and

Table 23.3 Bootstrapped success rates and rounding success rate lower bound for the ambiguity solutions of Table 23.2

Success rate	Original ambiguities	Transformed ambiguities
Lower bound rounding	0.51171	0.99995
Bootstrapping (1st ambiguity)	0.77749	0.99997
Bootstrapping (2nd ambiguity)	0.65816	0.99996

$\check{\mathbf{z}}_{\text{B}}^{(1)}$ are larger than those of their counterparts $\check{\mathbf{a}}_{\text{B}}^{(2)}$ and $\check{\mathbf{z}}_{\text{B}}^{(2)}$. This is due to the fact that in this example the first ambiguity is more precise than the second ambiguity. Thus bootstrapping should always start with the most precise ambiguity.

23.4 Integer Least-Squares

In this section we discuss the integer least-squares (ILS) ambiguity estimator. It has the best performance of all integer estimators. However, in contrast to rounding and bootstrapping, an integer search is needed for its computation.

23.4.1 Mixed Integer Least-Squares

Application of the least-squares principle to model (23.2), but now with the integer ambiguity constraints included, gives

$$(\check{\mathbf{a}}_{\text{LS}}, \check{\mathbf{b}}_{\text{LS}}) = \arg \min_{\mathbf{a} \in \mathbb{Z}^n, \mathbf{b} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{A}\mathbf{a} - \mathbf{B}\mathbf{b}\|_{\mathbf{Q}_{\text{yy}}}^2. \quad (23.50)$$

This is a nonstandard least-squares problem due to the integer constraints $\mathbf{a} \in \mathbb{Z}^n$ [23.14].

To solve (23.50), we start from the orthogonal decomposition

$$\|\mathbf{y} - \mathbf{A}\mathbf{a} - \mathbf{B}\mathbf{b}\|_{\mathbf{Q}_{\text{yy}}}^2 = \|\hat{\mathbf{e}}\|_{\mathbf{Q}_{\text{yy}}}^2 + \|\hat{\mathbf{a}} - \mathbf{a}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2 + \|\hat{\mathbf{b}}(\mathbf{a}) - \mathbf{b}\|_{\mathbf{Q}_{\hat{\mathbf{b}}(\mathbf{a})\hat{\mathbf{b}}(\mathbf{a})}}^2, \quad (23.51)$$

where $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{A}\hat{\mathbf{a}} - \mathbf{B}\hat{\mathbf{b}}$, with $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ the float solution, i. e., the unconstrained least-squares estimators of \mathbf{a} and \mathbf{b} respectively. Furthermore,

$$\hat{\mathbf{b}}(\mathbf{a}) = \hat{\mathbf{b}} - \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{a}}} \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}^{-1}(\hat{\mathbf{a}} - \mathbf{a}),$$

and

$$\mathbf{Q}_{\hat{\mathbf{b}}(\mathbf{a})\hat{\mathbf{b}}(\mathbf{a})} = \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{b}}} - \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{a}}} \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}^{-1} \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{b}}}.$$

Note that the first term on the right-hand side of (23.51) is constant and that the third term can be made zero for any \mathbf{a} by setting $\mathbf{b} = \hat{\mathbf{b}}(\mathbf{a})$. Hence, the mixed-integer minimizers of (23.50) are given as

$$\begin{aligned} \check{\mathbf{a}}_{\text{LS}} &= \arg \min_{\mathbf{z} \in \mathbb{Z}^n} \|\hat{\mathbf{a}} - \mathbf{z}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2, \\ \check{\mathbf{b}}_{\text{LS}} &= \hat{\mathbf{b}}(\check{\mathbf{a}}_{\text{LS}}) = \hat{\mathbf{b}} - \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{a}}} \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}^{-1}(\hat{\mathbf{a}} - \check{\mathbf{a}}_{\text{LS}}). \end{aligned} \quad (23.52)$$

In contrast to rounding and bootstrapping, the ILS principle is Z-invariant. For $\hat{\mathbf{z}} = \mathbf{Z}\hat{\mathbf{a}}$, we have

$$\check{\mathbf{z}}_{\text{LS}} = \mathbf{Z}\check{\mathbf{a}}_{\text{LS}} \quad \text{and} \quad \check{\mathbf{b}}_{\text{LS}} = \hat{\mathbf{b}} - \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{z}}} \mathbf{Q}_{\hat{\mathbf{z}}\hat{\mathbf{z}}}^{-1}(\hat{\mathbf{z}} - \check{\mathbf{z}}_{\text{LS}}). \quad (23.53)$$

Hence, application of the ILS principle to $\mathbf{Z}\hat{\mathbf{a}}$ gives the same result as \mathbf{Z} times the ILS estimator of \mathbf{a} . Also $\check{\mathbf{b}}_{\text{LS}}$ is invariant for the integer reparametrization.

The Z-invariance of the ILS principle also implies that the same success rate is obtained, i. e., $P(\check{\mathbf{z}}_{\text{LS}} = \mathbf{z}) = P(\check{\mathbf{a}}_{\text{LS}} = \mathbf{a})$. This is illustrated in Fig. 23.7. The number of green dots in the original scatterplot is exactly the same as the number of green dots in the transformed scatterplot.

When we compare Fig. 23.7 with Figs. 23.5 and 23.6, we note that the ILS pull-in region gives a better fit to the scatterplot than those of rounding and bootstrapping, thus indicating that ILS has a higher success rate. And indeed we have the following optimality property of the ILS estimator.

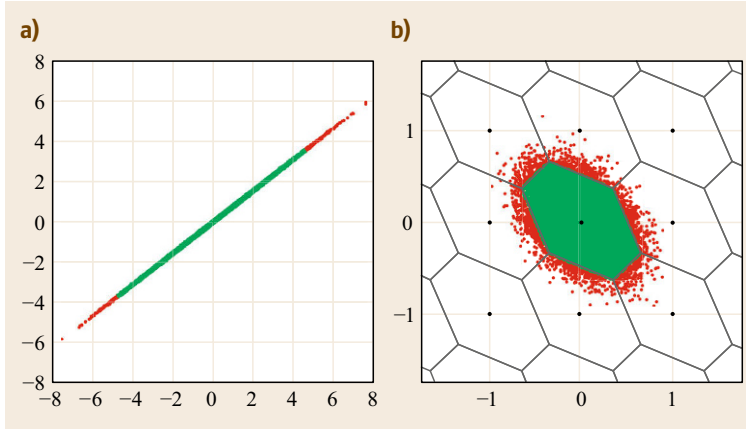


Fig. 23.7a,b 2-D ILS (original and transformed) pull-in regions and 50 000 float solutions. **(a)** Original ambiguities $\hat{\mathbf{a}}$ [cycles]; **(b)** Z-transformed ambiguities $\hat{\mathbf{z}} = \mathbf{Z}\hat{\mathbf{a}}$ [cycles]. Red dots will be pulled to wrong integer solutions, while green dots will be pulled to the correct integer solution

Theorem 23.4 ILS Optimality [23.35]

Let $\hat{\mathbf{a}} \sim N(\mathbf{a}, \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}})$. Then the integer least-squares estimator

$$\check{\mathbf{a}}_{\text{LS}} = \arg \min_{\mathbf{z} \in \mathbb{Z}^n} \|\hat{\mathbf{a}} - \mathbf{z}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2$$

has the largest success rate of all integer estimators. Furthermore

$$P(\check{\mathbf{a}}_{\text{R}} = \mathbf{a}) \leq P(\check{\mathbf{a}}_{\text{B}} = \mathbf{a}) \leq P(\check{\mathbf{a}}_{\text{LS}} = \mathbf{a}). \quad (23.54)$$

This result shows that there exists a clear ordering among the three most popular integer estimators. Integer rounding (IR) is the simplest, but it also has the poorest success rate. Integer least-squares (ILS) is the most complex, but also has the highest success rate of all. Integer bootstrapping (IB) sits in between. It does not need an integer search as is the case with ILS, and it does not completely neglect the information content of the ambiguity variance matrix as IR does.

The ordering (23.54) is illustrated by the empirical success rates in Table 23.4 for the cases shown in Figs. 23.5–23.7.

23.4.2 The ILS Computation

In this section the computation of the ILS solution (23.52) is presented. The two main parts of its computation are (a) the integer ambiguity search, and (b) the

ambiguity decorrelation. Although the ILS solution can in principle be computed on the basis of only (a), the decorrelation step is essential in the case of GNSS for improving the numerical efficiency of (a). This is particularly true in case of short observation time spans. Then the DD ambiguities turn out to be highly correlated due to the small change over time in the relative receiver–satellite geometry.

Integer Ambiguity Search

In contrast to rounding and bootstrapping, an integer search is needed to compute the ILS ambiguity solution

$$\check{\mathbf{a}} = \arg \min_{\mathbf{z} \in \mathbb{Z}^n} \|\hat{\mathbf{a}} - \mathbf{z}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2. \quad (23.55)$$

The search space is defined as

$$\Psi_{\mathbf{a}} = \{\mathbf{a} \in \mathbb{Z}^n \mid \|\hat{\mathbf{a}} - \mathbf{a}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2 \leq \chi^2\}, \quad (23.56)$$

where χ^2 is a to-be-chosen positive constant. This ellipsoidal search space is centered at $\hat{\mathbf{a}}$, its elongation is governed by $\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}$ and its size is determined by χ^2 . In the case of GNSS, the search space is usually extremely elongated due to the high correlations between the carrier-phase ambiguities. Since this extreme elongation hinders the computational efficiency of the search, the search space is first transformed to a more spherical shape by means of a decorrelating Z-transformation,

$$\Psi_{\mathbf{z}} = \{\mathbf{z} \in \mathbb{Z}^n \mid \|\hat{\mathbf{z}} - \mathbf{z}\|_{\mathbf{Q}_{\hat{\mathbf{z}}\hat{\mathbf{z}}}}^2 \leq \chi^2\}, \quad (23.57)$$

where $\hat{\mathbf{z}} = \mathbf{Z}\hat{\mathbf{a}}$ and $\mathbf{Q}_{\hat{\mathbf{z}}\hat{\mathbf{z}}} = \mathbf{Z}\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}\mathbf{Z}^T$.

In order for the search to be efficient, one would like the search space to be small such that it contains not too many integer vectors. This requires the choice of a small value for χ^2 , but one that still guarantees that the search space contains at least one integer vector. After all, $\Psi_{\mathbf{z}}$

Table 23.4 The percentages of correctly IR-, IB- and ILS-estimated ambiguities in original and transformed domain for the cases shown in Figs 23.5–23.7

Success rate	IR (%)	IB (%)	ILS (%)
Original $\hat{\mathbf{a}}$	23	29	97
Transformed $\hat{\mathbf{z}}$	95	96	97

has to be nonempty to guarantee that it contains the ILS solution \hat{z}_{LS} . Since the easy-to-compute (decorrelated) bootstrapped estimator gives a good approximation to the ILS estimator, \hat{z}_B is a good candidate for setting the size of the search space,

$$\chi^2 = \|\hat{z} - \hat{z}_B\|_{Q_{\hat{z}\hat{z}}}^2. \quad (23.58)$$

In this way one can work with a very small search space and still guarantee that the sought-for ILS solution is contained in it. If the rounding success rate is sufficiently high, one may also use \hat{z}_R instead of \hat{z}_B .

For the actual search, the quadratic form $\|\hat{z} - z\|_{Q_{\hat{z}\hat{z}}}^2$ is first written as a sum-of-squares. This is achieved by using the triangular decomposition $Q_{\hat{z}\hat{z}} = LDL^T$,

$$\|\hat{z} - z\|_{Q_{\hat{z}\hat{z}}}^2 = \sum_{i=1}^n \frac{(\hat{z}_{i|I} - z_i)^2}{\sigma_{i|I}^2} \leq \chi^2. \quad (23.59)$$

This sum-of-squares structure can now be used to set up the n intervals that are used for the search. These sequential intervals are given as

$$\begin{aligned} (\hat{z}_1 - z_1)^2 &\leq \sigma_1^2 \chi^2, \\ (\hat{z}_{2|1} - z_2)^2 &\leq \sigma_{2|1}^2 \left(\chi^2 - \frac{(\hat{z}_1 - z_1)^2}{\sigma_1^2} \right), \\ &\vdots \\ (\hat{z}_{n|(n-1), \dots, 1} - z_n)^2 &\leq \sigma_{n|(n-1), \dots, 1}^2 \\ &\quad \times \left(\chi^2 - \sum_{i=1}^{n-1} \frac{(\hat{z}_{i|I} - z_i)^2}{\sigma_{i|I}^2} \right). \end{aligned} \quad (23.60)$$

To search for all integer vectors that are contained in Ψ_z , one can now proceed as follows. First collect all integers z_1 that are contained in the first interval. Then for each of these integers, one computes the corresponding length and center point of the second interval, followed by collecting all integers z_2 that lie inside this second interval. By proceeding in this way to the last interval, one finally ends up with the set of integer vectors that lie inside Ψ_z . From this set one then picks the ILS solution as the integer vector that returns the smallest value for $\|\hat{z} - z\|_{Q_{\hat{z}\hat{z}}}^2$.

Various refinements on this search, with further efficiency improvements such as search space shrinking, are possible, see for example [23.27–29, 36, 37].

Ambiguity Decorrelation

To understand why the decorrelating Z-transformation is necessary to improve the efficiency of the search,

consider the structure of the sequential intervals (23.60) and assume that they are formulated for the original, nontransformed DD ambiguities of a single-baseline GNSS model. The DD ambiguity sequential conditional standard deviations $\sigma_{\hat{a}_{i|I}}$, $i = 1, \dots, n$, will then show a large discontinuity when going from the third to the fourth ambiguity.

As an example consider a single short baseline, with seven GPS satellites tracked, using dual-frequency phase-only data for two epochs separated by two seconds. Figure 23.8 shows its spectrum of sequential conditional standard deviations expressed in cycles, original as well as transformed. Note the logarithmic scale along the vertical axis. Since seven satellites were observed on both frequencies, we have twelve double-differenced ambiguities and therefore also twelve conditional standard deviations. The figure clearly shows the large drop in value when passing from the third to the fourth DD standard deviation, i.e., from $\sigma_{\hat{a}_{3|2,1}}$ to $\sigma_{\hat{a}_{4|3,2,1}}$. With the short time span, the DD ambiguities are namely poorly estimable, i.e., have large standard deviations, unless already three of them are assumed known, since with three DD ambiguities known, the baseline and remaining ambiguities can be estimated with a very high precision. Thus with $\sigma_{\hat{a}_1}$, $\sigma_{\hat{a}_{2|1}}$ and $\sigma_{\hat{a}_{3|2,1}}$ large, the first three bounds of (23.60), when formulated for the DD ambiguities, will be rather loose, while those of the remaining 9 inequalities will be very tight. As a consequence one will experience *search halting*. Of many of the collected integer candidates that satisfy the first three inequalities of (23.60), one will not

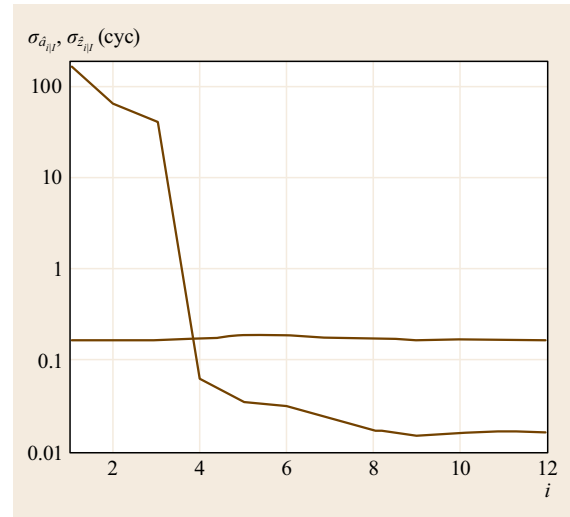


Fig. 23.8 Original and transformed (flattened) spectra of sequential conditional ambiguity standard deviations, $\sigma_{\hat{a}_{i|I}}$ and $\sigma_{\hat{z}_{i|I}}$, $i = 1, \dots, 12$, for a seven-satellite, dual-frequency, short GPS baseline (after [23.27])

be able to find corresponding integers that satisfy the remaining inequalities.

This inefficiency in the search is eliminated when using the Z-transformed ambiguities instead of the DD ambiguities. The decorrelating Z-transformation eliminates the discontinuity in the spectrum of sequential conditional standard deviations and, by virtue of the fact that the product of the sequential variances remains invariant (i. e., volume is preserved), also reduces the large values of the first three conditional variances.

In essence, the n -dimensional Z-transformation is constructed from two-dimensional decorrelating transformations as presented in Sect. 23.3.3. In two dimensions, the decorrelation achieves $\rho_z^2 \leq 1/4$ (23.46) and therefore

$$\sigma_{z_{2|1}}^2 = \sigma_{z_2}^2 (1 - \rho_z^2) \geq \frac{3}{4} \sigma_{z_2}^2 \geq \frac{3}{4} \sigma_{z_1}^2, \quad \text{if } \sigma_{z_1}^2 \leq \sigma_{z_2}^2.$$

Now let $\hat{a}_{i|I}$ and $\hat{a}_{i+1|I}$ play the role of \hat{a}_1 and \hat{a}_2 in the two-dimensional case. Then the decorrelation would achieve

$$\sigma_{z_{i+1|I+1}}^2 = \sigma_{z_{i+1|I}}^2 (1 - \rho_z^2) \geq \frac{3}{4} \sigma_{z_{i+1|I}}^2$$

and thus

$$\sigma_{z_{i+1|I+1}}^2 \geq \frac{3}{4} \sigma_{z_{i|I}}^2 \quad (23.61)$$

for $\sigma_{z_{i|I}}^2 \leq \sigma_{z_{i+1|I}}^2$. This shows that the originally large gap between $\sigma_{\hat{a}_{i|I}}$ and $\sigma_{\hat{a}_{i+1|I+1}}$, for $i = 3$, gets eliminated to a large extent, since now $\sigma_{z_{i+1|I+1}}$ cannot be much smaller than $\sigma_{z_{i|I}}$. Through a repeated application of such two-dimensional transformations, the whole spectrum of sequential conditional standard deviations can be flattened. In the case of Fig. 23.8 the transformed spectrum is flattened to a level slightly less than 0.2 cycles, while the original level for the DD standard deviations was more than 100 cycles.

The above described ILS procedure is mechanized in the GNSS LAMBDA (Least-squares AMBiguity Decorrelation Adjustment) method. For more information on the LAMBDA method, we refer to [23.14, 27–29, 37].

The following are examples for which one can see the LAMBDA method at work in a variety of different applications. Examples of such applications are baseline and network positioning [23.38–43], satellite formation flying [23.44–46], InSAR and VLBI [23.15, 16], GNSS attitude determination [23.47–50] and next-generation GNSS [23.51–53].

23.4.3 Least-Squares Success Rate

We have seen that the 2-D pull-in regions of rounding and bootstrapping are squares and parallelograms respectively. It follows that those of ILS are hexagons. The ILS pull-in region of $\mathbf{z} \in \mathbb{Z}^n$ consists by definition of all those points that are closer to \mathbf{z} than to any other integer vector in \mathbb{R}^n ,

$$\begin{aligned} \mathcal{L}_z &= \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{z}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2 \\ &\leq \|\mathbf{x} - \mathbf{u}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2, \forall \mathbf{u} \in \mathbb{Z}^n\}, \quad \mathbf{z} \in \mathbb{Z}^n. \end{aligned}$$

By rewriting the inequality, we obtain a representation that more closely resembles the ones of rounding \mathcal{R}_z and bootstrapping \mathcal{B}_z (23.21), (23.26),

$$\mathcal{L}_z = \left\{ \mathbf{x} \in \mathbb{R}^n \mid |w| \leq \frac{1}{2} \|\mathbf{u}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}, \quad \forall \mathbf{u} \in \mathbb{Z}^n \right\}, \quad (23.62)$$

with $\mathbf{z} \in \mathbb{Z}^n$ and

$$w = \frac{\mathbf{u}^\top \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}^{-1} (\mathbf{x} - \mathbf{z})}{\|\mathbf{u}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}} \quad (23.63)$$

the orthogonal projection of $(\mathbf{x} - \mathbf{z})$ onto the direction vector \mathbf{u} . This shows that the ILS pull-in regions are constructed from intersecting banded subsets centered at \mathbf{z} and having width $\|\mathbf{u}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}$. One can show that at most $2^n - 1$ of such subsets are needed for constructing the pull-in region. Note that $\mathcal{L}_z = \mathcal{R}_z$ when $\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}$ is diagonal.

The ILS PMF is given as

$$P(\check{\mathbf{a}}_{\text{LS}} = \mathbf{z}) = \int_{\mathcal{L}_z} \hat{f}_{\hat{\mathbf{a}}}(\mathbf{x}|\mathbf{a}) d\mathbf{x}. \quad (23.64)$$

To obtain the ILS success rate, set $\mathbf{z} = \mathbf{a}$.

Simulation

Due to the complicated geometry of the ILS pull-in regions, methods of Monte Carlo simulation are needed to evaluate the multivariate integral (23.64). Note that \mathbf{a} is not needed for the computation of the success rate. Thus one may simulate as if $\hat{\mathbf{a}}$ has the zero-mean distribution $N(\mathbf{0}, \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}})$. Also recall that the ILS success rate is Z-invariant, $P(\check{\mathbf{z}}_{\text{ILS}} = \mathbf{Z}\mathbf{a}) = P(\check{\mathbf{a}}_{\text{ILS}} = \mathbf{a})$. This property can be used to one's advantage when simulating. Since the simulation requires the repeated computation of an ILS solution, one is much better off doing this for a decorrelated $\hat{\mathbf{z}} = \mathbf{Z}\hat{\mathbf{a}}$, than for the original $\hat{\mathbf{a}}$.

The first step of the simulation is to use a random generator to generate n -independent samples from

the univariate standard normal distribution $N(0, 1)$, and then collect these in an n -vector \mathbf{s} . This vector is transformed as $\mathbf{G}\mathbf{s}$, with \mathbf{G} equal to the Cholesky factor of $\mathbf{Q}_{\hat{\mathbf{z}}\hat{\mathbf{z}}} = \mathbf{G}\mathbf{G}^\top$. The result is a sample $\mathbf{G}\mathbf{s}$ from $N(\mathbf{0}, \mathbf{Q}_{\hat{\mathbf{z}}\hat{\mathbf{z}}})$, and this sample is used as input for the ILS estimator. If the output of this estimator equals the null vector, then it is correct, otherwise it is incorrect. This simulation process can be repeated N number of times, and one can count how many times the null vector is obtained as a solution, say N_s times, and how often the outcome equals a nonzero integer vector, say N_f times. The approximations of the success rate and fail rate follow then as

$$P_s \approx \frac{N_s}{N} \quad \text{and} \quad P_f \approx \frac{N_f}{N}. \quad (23.65)$$

Further details on the success-rate simulation can be found in [23.18, 54, 55].

Lower and Upper Bounds

Instead of using simulation, one may also consider using bounds on the success rate. The following theorem

gives sharp lower and upper bounds on the ILS success rate.

Theorem 23.5 ILS success-rate bounds

Let $\hat{\mathbf{a}} \sim N(\mathbf{a}, \mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}})$, $\mathbf{a} \in \mathbb{Z}^n$, $\hat{\mathbf{z}} = \mathbf{Z}\hat{\mathbf{a}}$ and $c_n = (\frac{n}{2} \Gamma(\frac{n}{2}))^{2/n} / \pi$, with $\Gamma(\mathbf{x})$ the gamma function. Then

$$P(\check{\mathbf{z}}_B = \mathbf{z}) \leq P(\check{\mathbf{a}}_{\text{ILS}} = \mathbf{a}) \leq P\left(\chi_{n,0}^2 \leq \frac{c_n}{\text{ADOP}^2}\right) \quad (23.66)$$

for any admissible \mathbf{Z} -transformation and where $\chi_{n,0}^2$ denotes a random variable having a central chi-square distribution with n degrees of freedom.

The upper bound was first given in [23.56], albeit without proof. A proof is given in [23.24]. The lower bound was first given in [23.22, 35]. This lower bound (after decorrelation) is currently the sharpest lower bound available for the ILS success rate. A study on the performances of the various bounds can be found in [23.18, 54, 57, 58].

23.5 Partial Ambiguity Resolution

When the precision of the float ambiguity solution is poor, reliable integer estimation is not possible, i. e., the success rate will be too low. Instead of relying on the float solution and collecting more data, it might still be possible to reliably fix a subset of ambiguities, referred to as partial ambiguity resolution (PAR) [23.59].

The key issue is then the selection of the subset such that on the one hand the corresponding success rate will exceed a user-defined threshold, while at the same time it will result in a significant precision improvement of the position estimates. The first condition is important in order to prevent large positioning errors due to wrong fixing occurring. The second condition is optional, although it is obvious that PAR will only be beneficial if indeed the baseline precision is improved. Many options would be possible to select a subset of ambiguities to be fixed in the case of fixing the full set (FAR, fullset ambiguity resolution) is not possible or needed. Several approaches have been proposed in the literature in which it is first tried to fix only the (extra) widelane ambiguities in the case where two or more frequencies are being used [23.60–63]. Other ideas are to include only ambiguities with variances below a certain level, or ambiguities from satellites at a minimum elevation, with a minimum required signal-to-noise ratio, or which are visible for a certain time [23.59, 64].

Yet another strategy is to fix only (linear combinations of) ambiguities for which the best and second-best solutions are consistent [23.65]. A disadvantage of most of the PAR strategies is that the choice of the subset is not based on the success rate and/or precision improvement of the baseline solution. Moreover, some of the strategies involve an iterative procedure in which many different subsets are evaluated. This may require long search times.

The approach already proposed in [23.59] is easy to implement and does allow for choosing a minimum required success rate P_{\min} . The idea is to fix only the largest possible subset of decorrelated ambiguities, such that this success rate requirement can be met

$$\prod_{i=1}^k \left[2\Phi\left(\frac{1}{2\sigma_{z_{i|l}}}\right) - 1 \right] \geq P_{\min}. \quad (23.67)$$

Hence, only the first k entries of \mathbf{z} will be fixed, and the corresponding subset will be denoted as \mathbf{z}_S . Adding more ambiguities implies multiplication with another probability, which by definition is smaller than or equal to 1. Hence, k will be chosen such that the inequality in (23.67) holds, while a larger k (i. e., larger subset)

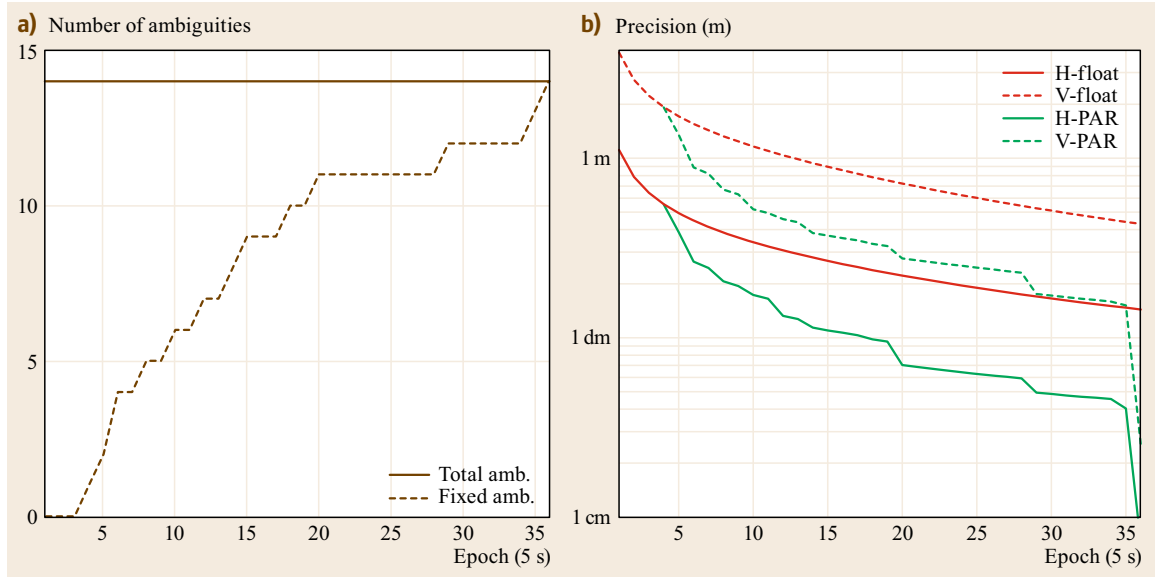


Fig. 23.9a,b Example of benefit of PAR for a 50 km baseline with a minimum required success rate of 99.9%. **(a)** Number of fixed ambiguities. **(b)** Baseline precision of float and partially fixed solutions

would result in a too low success rate. The corresponding precision improvement can be evaluated as well with

$$\mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{b}}} = \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{b}}} - \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{z}}_S} \mathbf{Q}_{\hat{\mathbf{z}}_S\hat{\mathbf{z}}_S}^{-1} \mathbf{Q}_{\hat{\mathbf{z}}_S\hat{\mathbf{b}}} \quad (23.68)$$

The uncertainty in the fixed subset solution can be ignored due to the high success rate requirement. An example of the benefit of PAR is shown in Fig. 23.9. It is an example of a dual-frequency 50 km baseline with eight GPS satellites tracked. The total number of ambiguities is thus equal to 14, and remains constant

for the whole timespan. Figure 23.9a shows the number of fixed ambiguities as function of the number of epochs based on recursive estimation. In this case full ambiguity resolution (FAR) is only possible after 36 epochs, but with PAR the number of fixed ambiguities gradually increases. For both PAR and FAR the minimum required success rate is set to 99.9%. The effect on the baseline precision is shown in the bottom panel. Both the precision of the vertical and horizontal baseline components start to improve with respect to the corresponding float precision as soon as a subset of the ambiguities is fixed.

23.6 When to Accept the Integer Solution?

So far no explicit description of the decision rule for accepting or rejecting the integer solution was given. In this section a flexible class of such rules is presented.

23.6.1 Model- and Data-Driven Rules

When do we accept the integer ambiguity solution $\check{\mathbf{a}}$? It was shown in Sect. 23.1.3 that working with the integer solution $\check{\mathbf{a}}$ only makes sense if the ambiguity success rate $P(\check{\mathbf{a}} = \mathbf{a})$ is sufficiently large or, equivalently, the fail rate $P(\check{\mathbf{a}} \neq \mathbf{a})$ is sufficiently small. Otherwise there would exist unacceptable chances of ending up with large errors in the fixed solution $\check{\mathbf{b}}$ (Fig. 23.3).

The above suggests the following decision rule for computing an outcome of the ambiguity resolution process,

$$\text{outcome} = \begin{cases} \check{\mathbf{a}} \in \mathbb{Z}^n & \text{if } P(\hat{\mathbf{a}} \notin \mathcal{P}_a) \leq P_0, \\ \hat{\mathbf{a}} \in \mathbb{R}^n & \text{otherwise.} \end{cases} \quad (23.69)$$

Thus with this rule the integer solution $\check{\mathbf{a}}$ is only accepted if the fail rate is smaller than a user-defined threshold P_0 . Otherwise it is rejected in favor of the float solution $\hat{\mathbf{a}}$. This is a *model-driven* rule, as the outcome is solely dependent on the strength of the underlying model. The actual data, i. e., the actual float solution $\hat{\mathbf{a}}$

itself, does not play a role in the decision. Only its PDF, through the probability $P(\hat{\mathbf{a}} \notin \mathcal{P}_a)$, affects the decision.

Instead of the model-driven rule (23.69), also a *data-driven* decision rule can be used. Such rules are of the form

$$\text{outcome} = \begin{cases} \check{\mathbf{a}} \in \mathbb{Z}^n & \text{if } \mathcal{T}(\hat{\mathbf{a}}) \leq \tau_0, \\ \hat{\mathbf{a}} \in \mathbb{R}^n & \text{otherwise,} \end{cases} \quad (23.70)$$

with testing function $\mathcal{T} : \mathbb{R}^n \mapsto \mathbb{R}$ and user-selected threshold value $\tau_0 \geq 0$. Thus in this case the integer solution $\check{\mathbf{a}}$ is accepted when $\mathcal{T}(\hat{\mathbf{a}})$ is sufficiently small; otherwise, it is rejected in favor of the float solution $\hat{\mathbf{a}}$. This rule is data driven, as the actual value of the float solution is used in the evaluation of $\mathcal{T}(\hat{\mathbf{a}})$.

In practice one usually uses a data-driven rule. Different choices for the testing function \mathcal{T} are then still possible. Examples include those of the ratio test, the difference test and the projector test. Each of these tests can be shown to be a member of the class of integer aperture (IA) estimators as introduced in [23.66–68]. A review and evaluation of these tests can be found in [23.54, 69–71].

The advantage of the data-driven rules over the model-driven rule (23.69) is the greater flexibility that they provide to the user, in particular with respect to the fail rate. With the data-driven rule, users can be given complete control over the fail rate, irrespective of the strength of the underlying GNSS model. This is impossible with the model-driven rule.

23.6.2 Four Ambiguity Resolution Steps

By including the test (23.70) into the ambiguity resolution process, its four steps become:

1. *Float solution*: Compute the float solution $\hat{\mathbf{a}} \in \mathbb{R}^n$ and $\hat{\mathbf{b}} \in \mathbb{R}^p$.
2. *Integer solution*: Choose an integer map $\mathcal{I} : \mathbb{R}^n \mapsto \mathbb{Z}^n$ and compute the integer solution $\check{\mathbf{a}} = \mathcal{I}(\hat{\mathbf{a}})$. Since the user has no real control over the success rate $P_s = P(\check{\mathbf{a}} = \mathbf{a})$, confidence cannot be assured if one relies solely on the outcome $\check{\mathbf{a}}$ of this second step. This is why the next step is needed. The role of the ambiguity acceptance test is namely to provide confidence in the integer outcomes of ambiguity resolution.
3. *Accept/reject integer solution*: Choose a testing function $\mathcal{T} : \mathbb{R}^n \mapsto \mathbb{R}$, with threshold τ_0 , and execute the test. Accept $\check{\mathbf{a}}$ if $\mathcal{T}(\hat{\mathbf{a}}) \leq \tau_0$, otherwise reject in favor of the float solution $\hat{\mathbf{a}}$.
4. *Fixed solution*: Compute the fixed solution $\check{\mathbf{b}}$ if the integer solution $\check{\mathbf{a}}$ is accepted, otherwise stick with the float solution $\hat{\mathbf{b}}$.

Due to the inclusion of the above ambiguity acceptance test, the quality of the outcome of the above four-step procedure will be different from that of the three-step procedure discussed in Sect. 23.1.2. We now determine the quality of the above four-step procedure.

23.6.3 Quality of Accepted Integer Solution

The integer $\check{\mathbf{a}} = \mathbf{z}$ is the outcome of the above step 3 (23.70) if both the conditions

$$\hat{\mathbf{a}} \in \mathcal{P}_z \text{ and } \mathcal{T}(\hat{\mathbf{a}}) \leq \tau_0 \quad (23.71)$$

are satisfied. Thus

$$\check{\mathbf{a}} = \mathbf{z} \text{ iff } \hat{\mathbf{a}} \in \Omega_z = \mathcal{P}_z \cap \Omega, \quad (23.72)$$

with acceptance region

$$\Omega = \{\mathbf{x} \in \mathbb{R}^n \mid \mathcal{T}(\mathbf{x}) \leq \tau_0\}. \quad (23.73)$$

The intersecting region $\Omega_z = \mathcal{P}_z \cap \Omega$ is called the *aperture pull-in region* of \mathbf{z} . The aperture pull-in regions are, just like the pull-in regions \mathcal{P}_z , translational invariant: $\Omega_z = \Omega_0 + \mathbf{z}$. The (green and red) ellipse-like regions of Fig. 23.10 are examples of such aperture pull-in regions. This figure also visualizes and summarizes which of the test outcomes are correct and which are not.

The outcome of the ambiguity acceptance test is correct if it is either the correct integer or a float solution that otherwise would be pulled to a wrong integer. The first happens when $\hat{\mathbf{a}} \in \Omega_a$, the second when $\hat{\mathbf{a}} \in \Omega^c \setminus (\mathcal{P}_a \setminus \Omega_a)$. The outcome is wrong if it is either the wrong integer or a float solution that otherwise would be pulled to the correct integer. The first happens when $\hat{\mathbf{a}} \in \Omega \setminus \Omega_a$, the second when $\hat{\mathbf{a}} \in \mathcal{P}_a \setminus \Omega_a$.

Once accepted by the test, the distribution of the integer $\check{\mathbf{a}}$ becomes a *conditional* PMF. Hence, instead of (23.12), we now have

$$P(\check{\mathbf{a}} = \mathbf{z} \mid \hat{\mathbf{a}} \in \Omega) = \frac{P(\hat{\mathbf{a}} \in \Omega_z)}{P(\hat{\mathbf{a}} \in \Omega)}. \quad (23.74)$$

Similarly, since the fixed solution is now only computed if $\check{\mathbf{a}}$ is accepted, its multimodal PDF is, instead of (23.14), given as

$$f_{\check{\mathbf{b}}}(\mathbf{x}) = \sum_{\mathbf{z} \in \mathbb{Z}^n} f_{\hat{\mathbf{b}}(\mathbf{z})}(\mathbf{x}) P(\check{\mathbf{a}} = \mathbf{z} \mid \hat{\mathbf{a}} \in \Omega). \quad (23.75)$$

As a wrong integer outcome, i. e., $\check{\mathbf{a}} \neq \mathbf{a}$, can result in large position errors (Fig. 23.3), it is of importance that sufficient confidence can be provided in the correctness of the integers as determined by the ambiguity

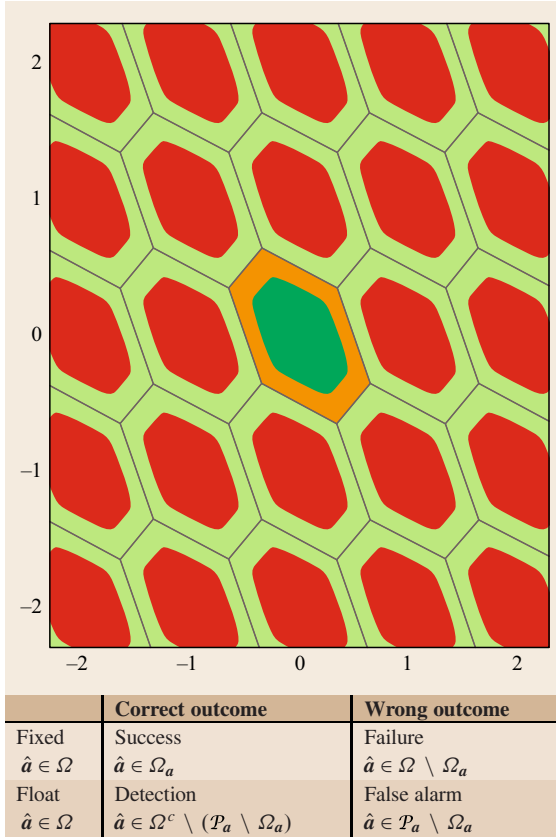


Fig. 23.10 Aperture pull-in regions $\Omega_z \subset \mathcal{P}_z$ and the four types of outcome: success (green), detection (light green), false alarm (orange) and failure (red) (after [23.71])

acceptance test. This confidence is described by the *probability of successful fixing*

$$P_{\text{SF}} = P(\check{\mathbf{a}} = \mathbf{a} | \hat{\mathbf{a}} \in \Omega) = \frac{P(\hat{\mathbf{a}} \in \Omega_a)}{P(\hat{\mathbf{a}} \in \Omega)}. \quad (23.76)$$

This is the conditional version of the unconditional success rate (23.13). It can be further expressed in the probability of *success*, $P_S = P(\hat{\mathbf{a}} \in \Omega_a)$, and the probability of *failure*, $P_F = P(\hat{\mathbf{a}} \in \Omega \setminus \Omega_a)$, as

$$P_{\text{SF}} = \frac{P_S}{P_S + P_F}. \quad (23.77)$$

From this important relation it follows that the user can now be given control over the probability of successful fixing. If, through an appropriate choice of the tolerance value τ_0 in (23.70), the aperture of Ω_0 is chosen to be sufficiently small, then $P_F \approx 0$ and therefore $P_{\text{SF}} \approx 1$, which, with (23.76) and (23.75), results in the peaked distribution $f_{\hat{\mathbf{b}}}(\mathbf{x}) \approx f_{\hat{\mathbf{b}}(a)}(\mathbf{x})$.

Thus with the inclusion of the ambiguity acceptance test, the user is given control over the quality of the integer outcome and thereby over the quality of the fixed solution $\hat{\mathbf{b}}$. This control is absent when only the three ambiguity resolution steps of Sect. 23.1.2 are used.

23.6.4 Fixed Failure-Rate Ratio Test

In practice, different testing functions \mathcal{T} are in use. Examples are those of the ratio test, the difference test or the projector test [23.12, 38, 72–76]. Here we describe the popular ratio test.

With the ratio test the ILS solution $\check{\mathbf{a}}$ is accepted iff

$$\mathcal{T}_R(\hat{\mathbf{a}}) = \frac{\|\hat{\mathbf{a}} - \check{\mathbf{a}}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2}{\|\hat{\mathbf{a}} - \check{\mathbf{a}}'\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2} \leq \tau_0, \quad (23.78)$$

with $0 < \tau_0 \leq 1$ and

$$\begin{aligned} \check{\mathbf{a}} &= \arg \min_{\mathbf{z} \in \mathbb{Z}^n} \|\hat{\mathbf{a}} - \mathbf{z}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2, \\ \check{\mathbf{a}}' &= \arg \min_{\mathbf{z} \in \mathbb{Z}^n, \mathbf{z} \neq \check{\mathbf{a}}} \|\hat{\mathbf{a}} - \mathbf{z}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2. \end{aligned} \quad (23.79)$$

The ratio test tests the closeness of the float solution to its nearest integer vector. If it is close enough, the test leads to acceptance of $\check{\mathbf{a}}$. If it is not close enough, then the test leads to rejection in favor of the float solution $\hat{\mathbf{a}}$.

The origin-centered aperture pull-in region of the ratio test is given as [23.70]

$$\begin{aligned} \Omega_{R,0} &= \left\{ \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2 \leq \tau_0 \|\mathbf{x} - \mathbf{z}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2 \right\} \\ &= \left\{ \mathbf{x} \in \mathbb{R}^n \mid \left\| \mathbf{x} + \frac{\tau_0}{1 - \tau_0} \mathbf{z} \right\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2 \leq \frac{\tau_0}{(1 - \tau_0)^2} \|\mathbf{z}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2 \right\} \end{aligned} \quad (23.80)$$

for all $\mathbf{z} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$. This shows that the aperture pull-in region is equal to the intersection of all ellipsoids with centers $-\tau_0/(1 - \tau_0)\mathbf{z}$ and radius $[\sqrt{\tau_0}/(1 - \tau_0)]\|\mathbf{z}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}$. Figure 23.11 shows two two-dimensional examples of the geometry of such aperture pull-in regions.

It is clear that the size or aperture of the pull-in region $\Omega_{R,0}$ determines the largest ratio \mathcal{T}_R one is willing to accept. The threshold value τ_0 can be used to tune this aperture. Smaller values corresponds to smaller apertures and thus smaller failure rates P_F . In the case where the threshold is taken equal to its maximal value $\tau_0 = 1$, the aperture pull-in regions become equal to the ILS pull-in regions, in which case the integer solution is always accepted. In such a case, the ratio test would be obsolete and can be discarded.

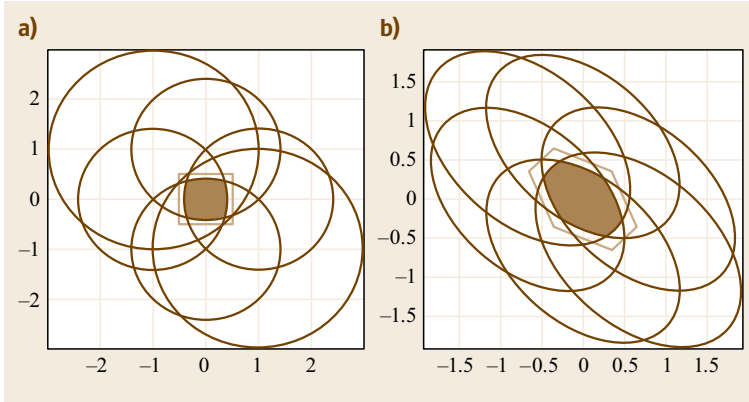


Fig. 23.11a,b Geometry of two-dimensional aperture pull-in region (*brown*) of the ratio test as constructed from intersecting circles (**a**) and ellipses (**b**) (after [23.70])

On the Choice of the Critical Value

The question is now how to choose the critical value τ_0 . Different values have been proposed in the literature, all based on empirical results. Typical values reported for τ_0 are $\frac{1}{3}$, $\frac{1}{2}$, and $\frac{2}{3}$ [23.3, 72, 75, 77]. The different values are already an indication that there is not one specific value that will always give the best performance. Care should therefore be exercised to consider these values generally applicable.

In [23.71, 78] it has been shown that the traditional usage of the ratio test, that is, with a fixed critical τ_0 -value, often results in either unacceptably high failure rates or is overly conservative. In the case of the next generation multifrequency, multi-GNSS models, for instance, the increase in strength of the models, due to, for example, more frequencies and more satellites, implies that the τ_0 -values can be chosen larger than the currently used fixed values. Thus for strong models, the fixed τ_0 -values currently in use are often too conservative, so that the false alarm rates are unnecessarily high, while the failure rates are very close to zero. For weak models, on the other hand, the currently used fixed τ_0 -values are often too large, so that the fixed solution is often wrongly accepted, thus resulting in high failure rates. These problems can be overcome if the ratio test is made adaptive to the strength of the underlying GNSS model.

It was therefore proposed in [23.67, 70, 71] to replace the fixed critical-value approach by the more flexible *fixed failure-rate* approach. With this approach, the user is given control over the failure rate for their particular application. Hence, depending on the requirements of the application (e.g., high, medium or low integrity), the user chooses a fixed value for the failure rate, say $P_F = 0.1\%$, and then computes the corresponding critical value τ_0 . The value of τ_0 will then adapt itself, in dependence on the underlying model strength, to ensure that the specified failure rate is indeed achieved (Fig. 23.12). In this way each project

or experiment can be executed with an a priori specified and guaranteed failure rate. The numerical procedure for computing τ_0 from P_F is described in [23.71] and is implemented in the LAMBDA-package (version 3).

23.6.5 Optimal Integer Ambiguity Test

As mentioned, the ratio test is not the only test with which the integer ambiguities can be validated. Hence, the fixed failure-rate approach can be applied to these other tests as well. Such work would then also be able to compare the performance of these tests and answer the question of which of the traditional tests, such as ratio test, difference test or projector test, performs best.

Instead of restricting attention to current tests, one can also take a more fundamental approach and try to determine an optimal test from first principles. This is the approach taken in [23.67, 68]. It resulted in the constrained maximum success-rate (CMS) test and the minimum mean penalty (MMP) test.

Constrained Maximum Success-Rate (CMS) Test

So far we considered fixed failure-rate ambiguity validation with an a priori given testing function \mathcal{T} or an a priori given aperture pull-in region Ω_0 . Instead of working with a predefined \mathcal{T} or Ω_0 , we now relax the situation and ask for which \mathcal{T} or Ω_0 the success rate P_S is maximized, given a user-defined failure rate P_F . The answer is given by the following theorem.

Theorem 23.6 Optimal integer ambiguity test [23.68]

Let $f_{\hat{\epsilon}}(\mathbf{x})$ and $f_{\hat{\epsilon}}(\mathbf{x})$ be the PDFs of the ambiguity residual vectors $\hat{\epsilon} = \hat{\mathbf{a}} - \hat{\mathbf{a}}$ and $\hat{\epsilon} = \hat{\mathbf{a}} - \mathbf{a}$ respectively. Then the solution to

$$\max_{\Omega_0} P_S \quad \text{subject to given } P_F \quad (23.81)$$

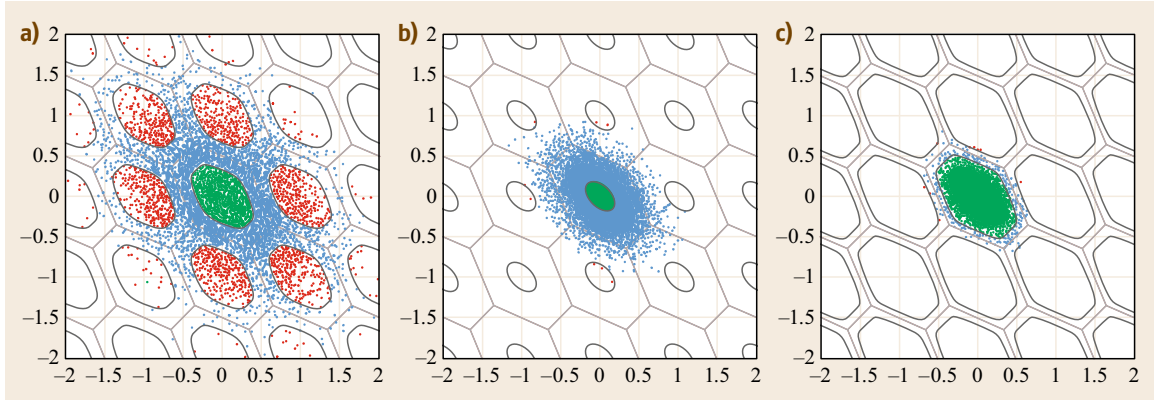


Fig. 23.12a–c A two-dimensional illustration of three different cases of integer ambiguity ratio-test validation. The *green* and *red dots* result in correct and incorrect integer outcomes respectively, while the *blue dots* result in the float solution as outcome. The first case **(a)** has poor performance, while the other two **(b,c)** have good performance. In the first case, due to an inappropriately chosen critical value τ_0 , the aperture pull-in region is too large thus producing too many wrong integer solutions. In the other two cases, the fixed failure-rate approach ($P_F = 0.1\%$) was used, thus resulting in critical values that adapt to the strength of the underlying model. As the second case **(b)** corresponds to a weaker model than the third case **(c)**, its aperture pull-in region is smaller thus producing more float solutions than in the third case. Both however have the same guaranteed small failure rate (after [23.79])

is given by the aperture pull-in region

$$\hat{\Omega}_0 = \left\{ \mathbf{x} \in \mathcal{P}_0 \mid \frac{f_{\hat{\epsilon}}(\mathbf{x})}{f_{\hat{\epsilon}}(\mathbf{x})} \geq \lambda \right\}, \quad (23.82)$$

with \mathcal{P}_0 the ILS pull-in region of the origin and λ ($0 < \lambda < 1$) the aperture parameter chosen so as to satisfy the a priori fixed failure rate P_F .

The PDFs of the ambiguity residuals are given as [23.80, 81],

$$\begin{aligned} f_{\hat{\epsilon}}(\mathbf{x}) &= \sum_{\mathbf{z} \in \mathbb{Z}^n} f_{\hat{\epsilon}}(\mathbf{x} + \mathbf{z}) p_0(\mathbf{x}), \\ f_{\hat{\epsilon}}(\mathbf{x}) &\propto \exp\left(-\frac{1}{2} \|\mathbf{x}\|_{\mathbf{Q}_{aa}}^2\right), \end{aligned} \quad (23.83)$$

with $p_0(\mathbf{x})$ the indicator function of \mathcal{P}_0 , i. e., $p_0(\mathbf{x}) = 1$ if $\mathbf{x} \in \mathcal{P}_0$ and $p_0(\mathbf{x}) = 0$ otherwise.

Note, since the PDFs $f_{\hat{\epsilon}}(\mathbf{x})$ and $f_{\hat{\epsilon}}(\mathbf{x})$ differ less, when $P(\hat{\mathbf{a}} = \mathbf{a}) \uparrow 1$, then the difference between the optimal aperture pull-in region $\hat{\Omega}_0$ and the ILS pull-in region \mathcal{P}_0 will also differ less when the ILS success rate increases. In the limit, all integer solutions will be accepted, since then $\hat{\Omega}_0 = \mathcal{P}_0$.

Minimum Mean Penalty (MMP) Test

This is also an optimal integer ambiguity acceptance test. The MMP test is based on the idea of penalizing certain outcomes of the test. The penalties, for example costs, are chosen by the user and can be made dependent on the application at hand. Different penalties are

assigned to different outcomes: a success penalty p_S if $\hat{\mathbf{a}} \in \Omega_a$ (green area in Fig. 23.10), a failure penalty p_F if $\hat{\mathbf{a}} \in \Omega \setminus \Omega_a$ (red area in Fig. 23.10), and a detection penalty p_D if $\hat{\mathbf{a}} \in \Omega^c \setminus (\mathcal{P}_a \setminus \Omega_a)$ (light green area in Fig. 23.10).

With this assignment, a discrete random variable, the penalty p , is constructed. It has three possible outcomes, $p = \{p_S, p_F, p_D\}$. We may now consider the average of the discrete random variable p , the average penalty $E(p)$, which is a weighted sum of the individual penalties, with the weights being equal to the three probabilities P_S , P_F , and P_D

$$E(p) = p_S P_S + p_F P_F + p_D P_D. \quad (23.84)$$

The MMP test is defined as the one having the smallest mean penalty. It follows from solving the minimization problem $\min_{\Omega_0} E(p)$ [23.67]. The solution is again given by (23.82), but now with the aperture parameter given as

$$\lambda = \frac{p_F - p_D}{p_F - p_S}. \quad (23.85)$$

Note that increasing the failure penalty p_F increases λ and contracts the aperture pull-in region $\hat{\Omega}_0$. This is as it should be, since a contracting $\hat{\Omega}_0$ reduces the occurrences of wrong integer solutions.

The Computational Steps

It is gratifying to see that the above two optimization principles provide the same structure for the optimal

ambiguity test. It implies, somewhat in analogy with the pairing of least-squares estimation and best linear unbiased estimation, that the same procedure can be given two different interpretations of optimality.

The steps for computing the CMS and MMP test are:

1. Compute the ILS ambiguity solution

$$\check{\mathbf{a}} = \arg \min_{\mathbf{z} \in \mathbb{Z}^n} \|\hat{\mathbf{a}} - \mathbf{z}\|_{\mathbf{Q}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}}^2. \quad (23.86)$$

2. Construct the ambiguity residual $\check{\epsilon} = \hat{\mathbf{a}} - \check{\mathbf{a}}$ and compute the PDF ratio

$$R(\check{\epsilon}) = \frac{f_{\check{\epsilon}}(\check{\epsilon})}{f_{\check{\epsilon}}(\check{\epsilon})}. \quad (23.87)$$

This outcome provides a measure of confidence in the solution. The larger the ratio, the more confidence one has. Note that the ratio can be seen as an approximation to the success fix-rate P_{SF} .

3. Determine the aperture parameter λ , either from the user-defined fail rate in the case of CMS, or from (23.85) in the case of MMP. Output the integer solution $\check{\mathbf{a}}$ if $R(\check{\epsilon}) \geq \lambda$, otherwise the outcome is the float solution $\hat{\mathbf{a}}$. Both $\check{\mathbf{a}}$ and $R(\check{\epsilon})$ can be computed efficiently with the LAMBDA method.

Acknowledgments. The author is the recipient of an Australian Research Council Federation Fellowship (project number FF0883188). This support is gratefully acknowledged.

References

- 23.1 G. Strang, K. Borre: *Linear Algebra, Geodesy, and GPS* (Wellesley Cambridge Press, Wellesley 1997)
- 23.2 P.J.G. Teunissen, A. Kleusberg (Eds.): *GPS for Geodesy*, 2nd edn. (Springer, Berlin 1998)
- 23.3 A. Leick, L. Rapoport, D. Tatarnikov: *GPS Satellite Surveying*, 4th edn. (Wiley, Hoboken 2015)
- 23.4 B. Hofmann-Wellenhof, H. Lichtenegger, E. Wasle: *GNSS Global Navigation Satellite Systems: GPS, GLONASS, Galileo & More* (Springer, Berlin 2007)
- 23.5 P. Misra, P. Enge: *Global Positioning System: Signals, Measurements, and Performance*, 2nd edn. (Ganga-Jamuna Press, Lincoln 2011)
- 23.6 P.J.G. Teunissen: Towards a unified theory of GNSS ambiguity resolution, *J. Global Position. Syst.* **2**(1), 1–12 (2003)
- 23.7 P.J.G. Teunissen: Mixed Integer Estimation and Validation for Next Generation GNSS. In: *Handbook of Geomathematics*, Vol. 2, ed. by W. Freeden, M.Z. Nashed, T. Sonar (Springer, Heidelberg 2010) pp. 1101–1127
- 23.8 C.C. Counselman, S.A. Gourevitch: Miniature interferometer terminals for earth surveying: ambiguity and multipath with Global Positioning System, *IEEE Trans. Geosci. Remote Sens.* **GE-19**(4), 244–252 (1981)
- 23.9 B.W. Remondi: Global Positioning System carrier phase: description and use, *J. Geodesy* **59**(4), 361–377 (1985)
- 23.10 R. Hatch: Dynamic differential GPS at the centimeter level, 4th Int. Geod. Symp. Satell. Position., Austin (Defense Mapping Agency / National Geodetic Survey, Silver Spring 1986) pp. 1287–1298
- 23.11 G. Blewitt: Carrier phase ambiguity resolution for the Global Positioning System applied to geodetic baselines up to 2000 km, *J. Geophys. Res.* **94**(B8), 10187–10203 (1989)
- 23.12 E. Frei, G. Beutler: Rapid static positioning based on the fast ambiguity resolution approach FARA: Theory and first results, *Manuscr. Geod.* **15**(6), 325–356 (1990)
- 23.13 J. Euler, H. Landau: Fast GPS ambiguity resolution on-the-fly for real-time applications, 6-th Int. Geodetic Symp. Satell. Posit., Columbus (Defense Mapping Agency, Springfield 1992) pp. 650–659
- 23.14 P.J.G. Teunissen: Least-squares estimation of the integer GPS ambiguities, Invited Lecture, Section IV Theory and Methodology, IAG General Meeting, Beijing (IAG, Budapest 1993)
- 23.15 T. Hobiger, M. Sekido, Y. Koyama, T. Kondo: Integer phase ambiguity estimation in next generation geodetic very long baseline interferometry, *Adv. Space Res.* **43**(1), 187–192 (2009)
- 23.16 B.M. Kampes, R.F. Hanssen: Ambiguity resolution for permanent scatterer interferometry, *IEEE Trans. Geosci. Remote Sens.* **42**(11), 2446–2453 (2004)
- 23.17 D. Catarino das Neves Viegas, S.R. Cunha: Precise positioning by phase processing of sound waves, *IEEE Trans. Signal Process* **55**(12), 5731–5738 (2007)
- 23.18 S. Verhagen, B. Li, P.J.G. Teunissen: Ps-LAMBDA: Ambiguity success rate evaluation software for interferometric applications, *Comput. Geosci.* **54**, 361–376 (2013)
- 23.19 P.J.G. Teunissen: On the integer normal distribution of the GPS ambiguities, *Artif. Satell.* **33**(2), 49–64 (1998)
- 23.20 C.P. Robert, G. Casella: *Monte Carlo Statistical Methods*, Vol. 2 (Springer, New York 1999)
- 23.21 P.J.G. Teunissen: The probability distribution of the GPS baseline for a class of integer ambiguity estimators, *J. Geod.* **73**(5), 275–284 (1999)
- 23.22 P.J.G. Teunissen: Success probability of integer GPS ambiguity rounding and bootstrapping, *J. Geod.* **72**(10), 606–612 (1998)
- 23.23 P.J.G. Teunissen: Influence of ambiguity precision on the success rate of GNSS integer ambiguity bootstrapping, *J. Geod.* **81**(5), 351–358 (2007)
- 23.24 P.J.G. Teunissen: ADOP based upper bounds for the bootstrapped and the least-squares ambiguity

- ity success-rates, *Artif. Satell.* **35**(4), 171–179 (2000)
- 23.25 P.J.G. Teunissen: The invertible GPS ambiguity transformations, *Manuscripta Geodaeica* **20**(6), 489–497 (1995)
- 23.26 P.J.G. Teunissen: A new method for fast carrier phase ambiguity estimation, *IEEE PLANS'94*, Las Vegas (IEEE, Piscataway 1994) pp. 562–573, doi:10.1109/PLANS.1994.303362
- 23.27 P.J.G. Teunissen: The least-squares ambiguity decorrelation adjustment: A method for fast GPS integer ambiguity estimation, *J. Geod.* **70**(1), 65–82 (1995)
- 23.28 P. De Jonge, C.C.J.M. Tiberius: The LAMBDA method for integer ambiguity estimation: Implementation aspects, *Publ. Delft Comput. Centre LGR-Ser.* **12**, 1–47 (1996)
- 23.29 P.J. De Jonge, C.C.J.M. Tiberius, P.J.G. Teunissen: Computational aspects of the LAMBDA method for GPS ambiguity resolution, *Proc. ION GPS 1996*, Kansas (ION, Virginia 1996) pp. 935–944
- 23.30 L.T. Liu, H.T. Hsu, Y.Z. Zhu, J.K. Ou: A new approach to GPS ambiguity decorrelation, *J. Geod.* **73**(9), 478–490 (1999)
- 23.31 E.W. Grafarend: Mixed integer-real valued adjustment (IRA) problems: GPS initial cycle ambiguity resolution by means of the LLL algorithm, *GPS Solut.* **4**(2), 31–44 (2000)
- 23.32 P. Xu: Random simulation and GPS decorrelation, *J. Geod.* **75**(7), 408–423 (2001)
- 23.33 P. Joosten, C. Tiberius: LAMBDA: FAQs, *GPS Solut.* **6**(1), 109–114 (2002)
- 23.34 J.G.G. Svendsen: Some properties of decorrelation techniques in the ambiguity space, *GPS Solut.* **10**(1), 40–44 (2006)
- 23.35 P.J.G. Teunissen: An optimality property of the integer least-squares estimator, *J. Geod.* **73**(11), 587–593 (1999)
- 23.36 P.J. de Jonge: A Processing Strategy for the Application of the GPS in Networks, Ph.D. Thesis (Delft University of Technology, Delft 1998)
- 23.37 X.W. Chang, X. Yang, T. Zhou: MLAMBDA: A modified LAMBDA method for integer least-squares estimation, *J. Geod.* **79**(9), 552–565 (2005)
- 23.38 C.C.J.M. Tiberius, P.J. De Jonge: Fast positioning using the LAMBDA method, 4th Int. Symp. on Differ. Satell. Navig. Syst. (DSNS'95), Bergen (1995) pp. 1–8
- 23.39 P.J. de Jonge, C.C.J.M. Tiberius: Integer estimation with the LAMBDA method, *IAG Symp. No. 115*, GPS Trends Terr. Airborne Spaceborne appl., Boulder, ed. by G. Beutler (Springer Verlag, Berlin 1996) pp. 280–284
- 23.40 F. Boon, B. Ambrosius: Results of real-time applications of the LAMBDA method in GPS based aircraft landings, *Int. Symp. on Kinemat. Syst. Geod., Geomat. Navig. (KIS'97)*, Banff (Dept. of Geomatics Engineering, Univ. of Calgary, Banff 1997) pp. 339–345
- 23.41 D. Odijk: Fast Precise GPS Positioning in the Presence of Ionospheric Delays, Ph.D. Thesis (Delft University of Technology, Delft 2002)
- 23.42 P.J.G. Teunissen, P.J. de Jonge, C.C.J.M. Tiberius: The least-squares ambiguity decorrelation adjustment: Its performance on short GPS baselines and short observation spans, *J. Geod.* **71**(10), 589–602 (1997)
- 23.43 Y.F. Tsai, J.C. Juang: Ambiguity resolution validation based on LAMBDA and eigen-decomposition, *Proc. ION AM 2007*, Cambridge (ION, Virginia 2001) pp. 299–304
- 23.44 D.B. Cox, J.D.W. Brading: Integration of LAMBDA ambiguity resolution with Kalman filter for relative navigation of spacecraft, *Navigation* **47**(3), 205–210 (2000)
- 23.45 S.C. Wu, Y.E. Bar-Sever: Real-time sub-cm differential orbit determination of two low-Earth orbiters with GPS bias-fixing, *Proc. ION GNSS 2006*, Fort Worth (ION, Virginia 2006) pp. 2515–2522
- 23.46 P.J. Buist, P.J.G. Teunissen, G. Giorgi, S. Verhagen: Instantaneous multi-baseline ambiguity resolution with constraints, *Int. Symp. on GPS/GNSS*, Tokyo, ed. by A. Yasuda (Japan Institute of Navigation, Tokyo 2008) pp. 862–871
- 23.47 C. Park, P.J.G. Teunissen: A new carrier phase ambiguity estimation for GNSS attitude determination systems, *Int. Symp. on GPS/GNSS*, Tokyo (2003) pp. 1–8
- 23.48 L. Dai, K.V. Ling, N. Nagarajan: Real-time attitude determination for microsatellite by LAMBDA method combined with Kalman filtering, *22nd AIAA Int. Commun. Satell. Syst. Conf. Exhib.*, Monterey (AIAA, Reston 2004) pp. 136–143
- 23.49 R. Monikes, J. Wendel, G.F. Trommer: A modified LAMBDA method for ambiguity resolution in the presence of position domain constraints, *Proc. ION GNSS 2005*, Long Beach (ION, Virginia 2005) pp. 81–87
- 23.50 G. Giorgi, P.J.G. Teunissen, P. Buist: A search and shrink approach for the baseline constrained LAMBDA method: Experimental results, *Int. Symp. on GPS/GNSS*, Tokyo, ed. by A. Yasuda (Japan Institute of Navigation, Tokyo 2008) pp. 797–806
- 23.51 B. Eissfeller, C.C.J.M. Tiberius, T. Pany, R. Biberger, T. Schueler, G. Heinrichs: Instantaneous ambiguity resolution for GPS/Galileo RTK positioning, *J. Gyrosc. Navig.* **38**(3), 71–91 (2002)
- 23.52 F. Wu, N. Kubo, A. Yasuda: Performance evaluation of GPS augmentation using quasi-zenith satellite system, *IEEE Trans. Aerosp. Electron. Syst.* **40**(4), 1249–1260 (2004)
- 23.53 S. Ji, W. Chen, C. Zhao, X. Ding, Y. Chen: Single epoch ambiguity resolution for Galileo with the CAR and LAMBDA methods, *GPS Solut.* **11**(4), 259–268 (2007)
- 23.54 S. Verhagen: The GNSS Integer Ambiguities: Estimation and Validation, Ph.D. Thesis (Delft University of Technology, Delft 2005)
- 23.55 S. Verhagen: On the reliability of integer ambiguity resolution, *Navigation* **52**(2), 99–110 (2005)
- 23.56 A. Hassibi, S. Boyd: Integer parameter estimation in linear models with applications to GPS, *IEEE Trans. Signal Process.* **46**(11), 2938–2952 (1998)
- 23.57 H.E. Thomsen: Evaluation of Upper and Lower Bounds on the Success Probability, *Proc. ION GPS 2000*, Salt Lake City (ION, Virginia 2000) pp. 183–188

- 23.58 S. Verhagen: On the approximation of the integer least-squares success rate: Which lower or upper bound to use?, *J. Global Position. Syst.* **2**(2), 117–124 (2003)
- 23.59 P.J.G. Teunissen, P. Joosten, C.C.J.M. Tiberius: Geometry-free ambiguity success rates in case of partial fixing, *Proc. ION NTM 1999*, San Diego (ION, Virginia 1999) pp. 201–207
- 23.60 B. Forsell, M. Martin Neira, R. Harris: Carrier phase ambiguity resolution in GNSS-2, *Proc. ION GPS 1997*, Kansas City (ION, Virginia 1997) pp. 1727–1736
- 23.61 R. Hatch, J. Jung, P. Enge, B. Pervan: Civilian GPS: the benefits of three frequencies, *GPS Solut.* **3**(4), 1–9 (2000)
- 23.62 P.J.G. Teunissen, P. Joosten, C.C.J.M. Tiberius: A comparison of TCAR, CIR and LAMBDA GNSS ambiguity resolution, *Proc. ION GPS 2002*, Portland (ION, Virginia 2002) pp. 2799–2808
- 23.63 B. Li, Y. Feng, Y. Shen: Three carrier ambiguity resolution: Distance-independent performance demonstrated using semi-generated triple frequency GPS signals, *GPS Solut.* **14**(2), 177–184 (2010)
- 23.64 A. Parkins: Increasing GNSS RTK availability with a new single-epoch batch partial ambiguity resolution algorithm, *GPS Solut.* **15**(4), 391–402 (2011)
- 23.65 D.G. Lawrence: A new method for partial ambiguity resolution, *Proc. ION ITM 2009*, Anaheim (ION, Virginia 2009) pp. 652–663
- 23.66 P.J.G. Teunissen: Integer aperture GNSS ambiguity resolution, *Artif. Satell.* **38**(3), 79–88 (2003)
- 23.67 P.J.G. Teunissen: Penalized GNSS Ambiguity Resolution, *J. Geodesy* **78**(4), 235–244 (2004)
- 23.68 P.J.G. Teunissen: GNSS ambiguity resolution with optimally controlled failure-rate, *Artif. Satell.* **40**(4), 219–227 (2005)
- 23.69 S. Verhagen: Integer ambiguity validation: An open problem?, *GPS Solut.* **8**(1), 36–43 (2004)
- 23.70 S. Verhagen, P.J.G. Teunissen: New global navigation satellite system ambiguity resolution method compared to existing approaches, *J. Guid. Control Dyn.* **29**(4), 981–991 (2006)
- 23.71 S. Verhagen, P.J.G. Teunissen: The ratio test for future GNSS ambiguity resolution, *GPS Solut.* **17**(4), 535–548 (2013)
- 23.72 H.J. Euler, B. Schaffrin: On a measure for the discernibility between different ambiguity solutions in the static kinematic GPS-mode, *IAG Symp. No. 107 Kinemat. Syst. Geodesy Surv. Remote Sens.*, Tokyo, ed. by I.I. Mueller (Springer Verlag, New York, Berlin, Heidelberg 1991) pp. 285–295
- 23.73 H. Landau, H.J. Euler: On-the-fly ambiguity resolution for precise differential positioning, *Proc. ION GPS 1992*, Albuquerque (ION, Virginia 1992) pp. 607–613
- 23.74 H.Z. Abidin: Computational and Geometrical Aspects of On-The-Fly Ambiguity Resolution, Ph.D. Thesis (University of New Brunswick, New Brunswick 1993)
- 23.75 S. Han: Quality-control issues relating to instantaneous ambiguity resolution for real-time GPS kinematic positioning, *J. Geodesy* **71**(6), 351–361 (1997)
- 23.76 J. Wang, M.P. Stewart, M. Tsakiri: A discrimination test procedure for ambiguity resolution on-the-fly, *J. Geodesy* **72**(11), 644–653 (1998)
- 23.77 M. Wei, K.P. Schwarz: Fast ambiguity resolution using an integer nonlinear programming method, *Proc. ION GPS 1995*, Palm Springs (ION, Virginia 1995) pp. 1101–1110
- 23.78 P.J.G. Teunissen, S. Verhagen: The GNSS ambiguity ratio-test revisited: A better way of using it, *Surv. Rev.* **41**(312), 138–151 (2009)
- 23.79 P.J.G. Teunissen, S. Verhagen: Integer aperture estimation: A framework for GNSS ambiguity acceptance testing, *Inside GNSS* **6**(2), 66–73 (2011)
- 23.80 P.J.G. Teunissen: The parameter distributions of the integer GPS model, *J. Geodesy* **76**(1), 41–48 (2002)
- 23.81 S. Verhagen, P.J.G. Teunissen: On the probability density function of the GNSS ambiguity residuals, *GPS Solut.* **10**(1), 21–28 (2006)

24. Batch and Recursive Model Validation

Peter J.G. Teunissen

Modeling errors, when passed unnoticed, may seriously deteriorate the final results of any estimation process. It is therefore of importance to have quality control procedures in place so as to be able to judge and validate the outcome of estimation. This chapter presents such methods for global navigation satellite system (GNSS) model validation and qualification. Since batch and recursive estimation are common methods of GNSS data processing, the validation and integrity monitoring of both will be discussed in this chapter.

24.1	Modeling and Validation	687
24.2	Batch Model Validation	689
24.2.1	Null versus Alternative Hypothesis	689
24.2.2	Unbiased versus Biased Solution	689
24.2.3	Effect of the Influential Bias	690
24.3	Testing for a Bias	692
24.3.1	The Most Powerful Test Statistic	692
24.3.2	Alternative Expressions for Test Statistic T_q	695
24.3.3	Test Statistic T_q Expressed in LS Residuals	696
24.3.4	Optimality of the w -Test Statistic	697
24.3.5	The Minimal Detectable Bias	697
24.3.6	Hazardous Missed Detection	700
24.4	Testing Procedure	705
24.4.1	Detection, Identification and Adaptation	705
24.4.2	Data Snooping	706
24.4.3	Unknowns in the Stochastic Model	709
24.5	Recursive Model Validation	710
24.5.1	Model and Filter	710
24.5.2	Models and UMPI Test Statistic	711
24.5.3	Local and Global Testing	711
24.5.4	Recursive Detection	712
24.5.5	Recursive Identification	713
24.5.6	Recursive Adaptation: General Case	714
24.5.7	Recursive Adaptation: Special GNSS Case	715
	References	717

24.1 Modeling and Validation

When linking a mathematical model to data, care has to be exercised in formulating its functional and stochastic model. The functional model describes the relations that are believed to exist between the observables and the unknown parameters, while the stochastic model is used to capture the expected uncertainty or variability of the data.

We write

$$\mathbf{y} \sim \mathcal{N}(E(\mathbf{y}), D(\mathbf{y})) , \quad (24.1)$$

to denote that the data vector \mathbf{y} is assumed to be distributed as (\sim) a normal distribution (\mathcal{N}) with expectation $E(\mathbf{y})$ and dispersion $D(\mathbf{y})$.

In the case of GNSSs, the functional model, $E(\mathbf{y}) = \mathbf{A}(\mathbf{x})$, links the pseudorange (code) and/or carrier-phase observables collected in vector $\mathbf{y} \in \mathbb{R}^m$ to the

unknown entries of the parameter vector $\mathbf{x} \in \mathbb{R}^n$, for example coordinates and clocks (receiver, satellite), atmospheric delays (troposphere, ionosphere), hardware/environmental delays (instrumental biases, multipath), and carrier-phase ambiguities (Chap. 21). Whether or not all these parameters need to be included depends very much on the circumstances of measurement and the particular application at hand. For instance, differential ionospheric delays could be negligible in the case of sufficiently short baselines, or multipath may be avoided when sufficient precautions in the measurement setup are taken.

In the stochastic model, $D(\mathbf{y}) = \mathbf{Q}_{yy}$, one tries to capture the intrinsic uncertainties of the measurements. For this one needs a proper understanding of the instrumentation and the measurement procedures used. Formulating the stochastic model does not only involve

the precision of the individual measurements, but may also involve cross correlation and/or time correlation. Depending on how the measurement process is implemented in the GNSS receivers, the observables may or may not be cross correlated (Chaps. 13 and 14). Also the potential presence of time correlation should be considered, in particular when use is made of high sampling rates.

Depending on the application at hand, one can choose from a whole suite of GNSS functional models $\mathbf{A} : \mathbb{R}^n \mapsto \mathbb{R}^m$ (Chap. 21). A GNSS model for relative positioning may be based on the simultaneous use of two receivers (single-baseline) or more than two receivers (multibaseline or network). It may have the relative receiver-satellite geometry included (geometry-based) or excluded (geometry-free). When it is excluded, it is not the baseline components that are involved as unknowns in the model, but instead the receiver-satellite ranges themselves. GNSS models may also be discriminated as to whether the receiver(s) are in motion (nonstationary) or not (stationary). When in motion, one solves for one or more trajectories, since with the receiver-satellite geometry included, one will have a new position for each new epoch.

It will be clear that various modeling errors can be made when formulating a mathematical model like

$$\mathbf{y} \sim \mathcal{N}(\mathbf{A}(\mathbf{x}), \mathbf{Q}_{yy}) . \quad (24.2)$$

The functional model could be misspecified, $E(\mathbf{y}) \neq \mathbf{A}(\mathbf{x})$. The variance matrix could be misspecified $D(\mathbf{y}) \neq \mathbf{Q}_{yy}$, or even the assumed normal distribution of \mathbf{y} could be a poor approximation.

In this chapter we restrict attention to misspecifications in the mean only, as these are by far the most common modeling errors to occur in practice. Examples are cycle slips in phase data, outliers in code data, antenna-height errors, erroneous negligence of atmospheric delays, or any other underparametrization of the GNSS model, see for example [24.1–9]. As such modeling errors, when passed unnoticed, may seriously deteriorate the final results of the GNSS estimation process, it is of importance that proper quality control procedures of data and model are in place so as to be able to judge and validate the outcome of estimation.

This chapter presents such tools for GNSS model validation and qualification. The material presented is generally valid and therefore not dependent on any spe-

cific GNSS application. As it applies to the whole suite of different GNSS models, the quality control can be exercised at various different processing stages, such as at the channel/receiver/baseline/network levels. At a single-channel level for instance, it is already possible to validate the time series of undifferenced data. By combining some of the parameters (e.g., range, clock errors, tropospheric delay, orbital uncertainty) and by introducing a dynamic model on the time behavior of the ionospheric delays, the necessary redundancy enters that is needed for validation. Such single-channel data validation is useful for detecting the larger anomalies at an early stage; see for example [24.10–13].

Due to the additional redundancy, more powerful testing becomes possible at the receiver and baseline level [24.14–19]. In this case the observation equations are parametrized in terms of the position- or baseline coordinates of a single (e.g., in case of precise point positioning (PPP)) or dual (e.g., in case of real time kinematic (RTK)) receiver. Here the redundancy primarily stems from the presence of the receiver-satellite geometry in the design matrix and from the assumed time constancy of the ambiguities. Additional redundancy enters when the receiver position is considered stationary instead of moving, and/or when an array or network of receivers is considered; see for example [24.20–22].

Since batch and recursive estimation are both common methods of GNSS data processing, the quality control of both will be discussed in this chapter. We start with the batch approach as this for a large part already paves the way for the recursive treatment. In Sect. 24.2 we formulate the hypothesis framework for testing and introduce the concepts of testable and influential biases. The general form of the most powerful test statistic is introduced in Sect. 24.3, together with a discussion of its properties and its usage for GNSS. This includes the important concepts of minimal detectable biases (MDBs) and hazardous missed detection (HMD). A testing procedure for the handling of multiple hypotheses is discussed in Sect. 24.4. It includes separate steps for the detection, identification and adaptation of modeling errors. This is extended in Sect. 24.5 to a Kalman-filter-based recursive algorithm for the detection, identification and adaptation of GNSS modeling errors. Depending on the required power of testing, a difference is hereby made between local and global testing. Various GNSS examples are given throughout the chapter to illustrate the concepts discussed.

24.2 Batch Model Validation

24.2.1 Null versus Alternative Hypothesis

Before any start can be made with statistical model validation, one needs to have a clear idea of the null and alternative hypotheses: \mathcal{H}_0 and \mathcal{H}_a respectively. The null hypothesis, also referred to as working hypothesis, consists of the model that one believes to be valid under normal working conditions. We assume the null hypothesis to be of the form

$$\mathcal{H}_0: E(\mathbf{y}) = \mathbf{A}\mathbf{x}, D(\mathbf{y}) = \mathbf{Q}_{yy}, \quad (24.3)$$

with vector of observables $\mathbf{y} \in \mathbb{R}^m$, vector of unknown parameters $\mathbf{x} \in \mathbb{R}^n$, known $m \times n$ design matrix \mathbf{A} of full rank $\text{rank}(\mathbf{A}) = n$, and known positive-definite $m \times m$ variance-matrix \mathbf{Q}_{yy} . In case of GNSS, the model (24.3) is formed by the linear(ized) code and/or carrier-phase observation equations; Chaps. 19, 21, and [24.5, 7, 23, 24].

In order to test the null hypothesis against an alternative hypothesis, we need to know what kind of misspecifications one can expect. Although every part of the null hypothesis can be wrong of course, we assume here that the misspecifications are confined to mismodeling of the mean of \mathbf{y} , $E(\mathbf{y})$. Experience has shown that these are by and large the most common errors that occur when formulating the model. Hence, the alternative hypothesis takes the form

$$\mathcal{H}_a: E(\mathbf{y}) = \mathbf{A}\mathbf{x} + \mathbf{C}\mathbf{b}, D(\mathbf{y}) = \mathbf{Q}_{yy}, \quad (24.4)$$

with $[\mathbf{A}, \mathbf{C}]$ a known full-rank matrix of order $m \times (n+q)$ and $\mathbf{b} \in \mathbb{R}^q \setminus \{0\}$ the additional unknown parameter vector. The number of entries in \mathbf{b} , i.e., q , can range from 1 to $m-n$. Thus the maximum number of additional parameters that can be accommodated equals the redundancy under \mathcal{H}_0 . In that case $q = m-n$ and $[\mathbf{A}, \mathbf{C}]$ becomes a square and invertible matrix, implying that no restrictions are put anymore on the mean of \mathbf{y} , $E(\mathbf{y}) \in \mathbb{R}^m$.

The vector

$$\mathbf{b}_y = \mathbf{C}\mathbf{b} \quad (24.5)$$

in (24.4) models the bias that is potentially present in the mean of \mathbf{y} . For instance, through $\mathbf{b}_y = \mathbf{C}\mathbf{b}$ one may model the presence of one or more blunders (outliers) in the data, or the presence of neglected atmospheric effects, or any other systematic effect that one failed to take into account under \mathcal{H}_0 .

Here we give some such GNSS examples of the $m \times q$ matrix \mathbf{C} (note: in the case $q = 1$, the $m \times q$ matrix

\mathbf{C} becomes a vector, which we denote with the lower case \mathbf{c}):

Outlier in i th code observable: if the vector of observables \mathbf{y} consists of code data only, then $q = 1$ and $\mathbf{c} = (0, \dots, 0, 1, 0, \dots, 0)^T$, with the 1 as the i th entry of \mathbf{c} .

Cycle-slip in carrier-phase that started at epoch $l \leq k$: if \mathbf{y} consists of single-channel carrier-phase data covering k epochs, then $q = 1$ and $\mathbf{c} = (0, \dots, 0, 1, 1, \dots, 1)^T$, with the 1s occupying the last $k-l+1$ entries of \mathbf{c} .

Baseline error in i th baseline of a GNSS baseline network: if \mathbf{y} consists of the baseline vectors, then $q = 3$ and $\mathbf{C} = (0, \dots, 0, \mathbf{I}_3, 0, \dots, 0)^T$, with the unit matrix \mathbf{I}_3 as the i th matrix entry.

Antenna height error at GNSS receiver location: if \mathbf{y} consists of Cartesian baseline vectors, then $q = 1$ and $\mathbf{c} = (0, \dots, 0, \partial N/\partial h, \partial E/\partial h, \partial U/\partial h, 0, \dots, 0)^T$, with the nonzero entries being the elements of the linearized transformation matrix from N-E-U Cartesian to ellipsoidal coordinates.

Satellite failure Let $\mathbf{y} = (\dots, \mathbf{y}^s, \dots)^T$ be the vector of observables, with \mathbf{y}^s containing all m_s observations to the failed satellite s . Then $q = m_s$ and $\mathbf{C} = (0, \dots, 0, \mathbf{I}_{m_s}, 0, \dots, 0)^T$, under the assumption that a failed satellite affects all m_s observations to that satellite s .

Ionospheric gradient in single-frequency array signals to pivot-satellite s . Let \mathbf{y} consist of the $2(m-1)(n-1)$ single-epoch double-differenced (DD) code and phase data of a small three-dimensional array of n receivers, tracking m satellites of which the pivot-satellite data are assumed affected by the ionospheric gradient. Then $q = 3$, \mathbf{b} is the three-dimensional ionospheric gradient, and

$$\mathbf{C} = [(\mathbf{e} \otimes \mathbf{B}^T)^T, -(\mathbf{e} \otimes \mathbf{B}^T)^T]^T,$$

with \mathbf{e} the $(m-1)$ -vector of 1s, \otimes the Kronecker product, and $\mathbf{B} = [\mathbf{x}_1, \dots, \mathbf{x}_{n-1}]$, the $3 \times (n-1)$ matrix of array baseline vectors.

These are just a few examples of how modeling errors can be cast in the formulation of (24.4) and (24.5). For any such case the theory and methods of this chapter apply. More examples can be found in [24.20, 25–28].

24.2.2 Unbiased versus Biased Solution

Before discussing the testing of \mathcal{H}_0 against \mathcal{H}_a , we first consider the impact that a bias can have on our solutions. Under the working hypothesis \mathcal{H}_0 (24.3), we would compute the least-squares (LS) estimators of \mathbf{x}

and \mathbf{Ax} , together with the residual vector, as

$$\begin{aligned}\hat{\mathbf{x}} &= \mathbf{A}^+ \mathbf{y}, \\ \hat{\mathbf{y}} &= \mathbf{P}_A \mathbf{y}, \\ \hat{\mathbf{e}} &= \mathbf{P}_A^\perp \mathbf{y},\end{aligned}\quad (24.6)$$

with LS inverse

$$\mathbf{A}^+ = (\mathbf{A}^T \mathbf{Q}_{yy}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Q}_{yy}^{-1}$$

and orthogonal projectors

$$\mathbf{P}_A = \mathbf{A} \mathbf{A}^+ \quad \text{and} \quad \mathbf{P}_A^\perp = \mathbf{I}_m - \mathbf{P}_A.$$

The geometry of this LS estimation, with its orthogonal decomposition $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$, is depicted in Fig. 24.1. Orthogonality is here defined in the metric defined by the variance matrix \mathbf{Q}_{yy} . One important consequence of the orthogonality $\hat{\mathbf{y}} \perp \hat{\mathbf{e}}$ is that the normally distributed vectors $\hat{\mathbf{y}}$ and $\hat{\mathbf{e}}$ are independent. Thus also $\hat{\mathbf{x}}$ is independent of $\hat{\mathbf{e}}$.

Under \mathcal{H}_0 and \mathcal{H}_a , the estimators $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{e}}$ are distributed as

$$\begin{aligned}\hat{\mathbf{x}} &\stackrel{\mathcal{H}_0}{\sim} \mathcal{N}(\mathbf{x}, \mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}), & \hat{\mathbf{x}} &\stackrel{\mathcal{H}_a}{\sim} \mathcal{N}(\mathbf{x} + \mathbf{b}_{\hat{\mathbf{x}}}, \mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}), \\ \hat{\mathbf{y}} &\stackrel{\mathcal{H}_0}{\sim} \mathcal{N}(\mathbf{Ax}, \mathbf{Q}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}), & \hat{\mathbf{y}} &\stackrel{\mathcal{H}_a}{\sim} \mathcal{N}(\mathbf{Ax} + \mathbf{b}_{\hat{\mathbf{y}}}, \mathbf{Q}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}), \\ \hat{\mathbf{e}} &\stackrel{\mathcal{H}_0}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}}), & \hat{\mathbf{e}} &\stackrel{\mathcal{H}_a}{\sim} \mathcal{N}(\mathbf{b}_{\hat{\mathbf{e}}}, \mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}}),\end{aligned}\quad (24.7)$$

with the variance matrices $\mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} = (\mathbf{A}^T \mathbf{Q}_{yy}^{-1} \mathbf{A})^{-1}$, $\mathbf{Q}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \mathbf{A} \mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} \mathbf{A}^T$, and $\mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}} = \mathbf{Q}_{yy} - \mathbf{Q}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ respectively, and the bias vectors given as

$$\begin{aligned}\mathbf{b}_{\hat{\mathbf{x}}} &= \mathbf{A}^+ \mathbf{b}_y, \\ \mathbf{b}_{\hat{\mathbf{y}}} &= \mathbf{P}_A \mathbf{b}_y, \\ \mathbf{b}_{\hat{\mathbf{e}}} &= \mathbf{P}_A^\perp \mathbf{b}_y.\end{aligned}\quad (24.8)$$

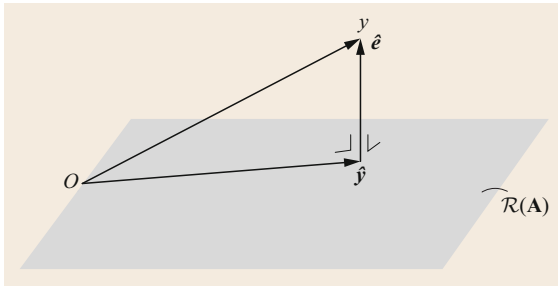


Fig. 24.1 Orthogonal decomposition of the vector of observables $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$, with $\hat{\mathbf{y}} = \mathbf{P}_A \mathbf{y} \in \mathcal{R}(\mathbf{A})$ and $\hat{\mathbf{e}} = \mathbf{P}_A^\perp \mathbf{y} \in \mathcal{R}(\mathbf{A})^\perp$

Hence, the LS estimators (24.6) are unbiased under the null hypothesis, but not so under the alternative hypothesis. Their biases (24.8) follow from propagating the bias \mathbf{b}_y (24.5) through (24.6).

A geometric interpretation similar to (24.6) can be given to (24.8) (Fig. 24.2). In analogy to $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$, we have the orthogonal decomposition

$$\mathbf{b}_y = \mathbf{b}_{\hat{\mathbf{y}}} + \mathbf{b}_{\hat{\mathbf{e}}} \quad (24.9)$$

in which we call

$$\begin{aligned}\mathbf{b}_y &= \text{actual bias}, \\ \mathbf{b}_{\hat{\mathbf{y}}} &= \text{influential bias}, \\ \mathbf{b}_{\hat{\mathbf{e}}} &= \text{testable bias}.\end{aligned}\quad (24.10)$$

As the vector $\mathbf{b}_{\hat{\mathbf{e}}}$ is expected to be zero under \mathcal{H}_0 (24.7), it is this component of the bias vector \mathbf{b}_y that can be tested. The component $\mathbf{b}_{\hat{\mathbf{y}}}$ however cannot be tested. It lies in the range space of matrix \mathbf{A} , $\mathbf{b}_{\hat{\mathbf{y}}} \in \mathcal{R}(\mathbf{A})$, and will therefore be directly absorbed by the parameter vector. Hence, this is the component of the bias vector \mathbf{b}_y that directly influences the parameter solution.

It will be clear that $\mathbf{b}_y \in \mathcal{R}(\mathbf{A})^\perp$ would be the most favorable situation. In that case the complete bias vector is testable, while any influence on the parameter solution would be absent. The worst situation occurs if $\mathbf{b}_y \in \mathcal{R}(\mathbf{A})$. Testability of the bias is then impossible and the complete bias vector will propagate into the parameter solution.

24.2.3 Effect of the Influential Bias

In order to evaluate the bias influence in a probabilistic sense, we consider the effect it has on the *confidence region* of the parameter estimator $\hat{\mathbf{x}}$. Since

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_{\mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}}^2 \stackrel{\mathcal{H}_0}{\sim} \chi^2(n, 0),$$

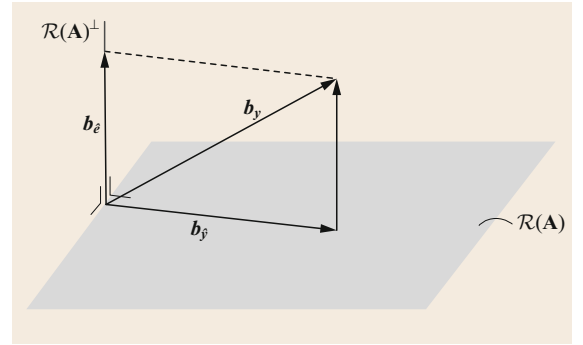


Fig. 24.2 Bias decomposition: orthogonal decomposition of actual bias $\mathbf{b}_y = \mathbf{b}_{\hat{\mathbf{y}}} + \mathbf{b}_{\hat{\mathbf{e}}}$ into influential bias $\mathbf{b}_{\hat{\mathbf{y}}} \in \mathcal{R}(\mathbf{A})$ and testable bias $\mathbf{b}_{\hat{\mathbf{e}}} \in \mathcal{R}(\mathbf{A})^\perp$ (after [24.29])

with

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_{\mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}}^2 = (\hat{\mathbf{x}} - \mathbf{x})^T \mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}^{-1} (\hat{\mathbf{x}} - \mathbf{x}),$$

the $100(1 - \eta)\%$ confidence region is given as

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_{\mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}}^2 \leq \chi_{\eta}^2(n, 0) \quad (24.11)$$

in which $\chi_{\eta}^2(n, 0)$ is the ordinate value of the $\chi^2(n, 0)$ -distribution above which we find an area of size η . Hence, under \mathcal{H}_0 , the probability that $\hat{\mathbf{x}}$ lies *outside* the \mathbf{x} -centered ellipsoidal region (24.11) is equal to $100\eta\%$.

In practice it is the criteria for system design that will provide information on what is considered an acceptable size and shape of the $100(1 - \eta)\%$ confidence region. Once η is given, one would know how precise the estimator $\hat{\mathbf{x}}$ needs to be in order to meet the prescribed geometry of the confidence region. Hence, one would then know what variance matrix $\mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} = (\mathbf{A}^T \mathbf{Q}_{yy}^{-1} \mathbf{A})^{-1}$ to aim for in the design of the system or design of the measurement scenario (i. e., through choice of \mathbf{A} and \mathbf{Q}_{yy}).

Since one would like the occurrence of $\hat{\mathbf{x}}$ falling outside the confidence region to be rare, the size η will usually be small. In safety-critical situations one would even consider such occurrences as *hazardous*. Such hazardous occurrences may not be rare however, if biases are present in the solution $\hat{\mathbf{x}}$. In the presence of biases, i. e., under \mathcal{H}_a , the probability of a hazardous occurrence will be larger than $100\eta\%$. To evaluate such a probability we make use of the noncentral χ^2 -distribution. We have

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_{\mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}}^2 \stackrel{\mathcal{H}_a}{\sim} \chi^2(n, \lambda_{\hat{\mathbf{x}}}^2), \quad (24.12)$$

with the noncentrality parameter given as

$$\lambda_{\hat{\mathbf{x}}}^2 = \mathbf{b}_{\hat{\mathbf{x}}}^T \mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}^{-1} \mathbf{b}_{\hat{\mathbf{x}}} = \mathbf{b}_{\hat{\mathbf{y}}}^T \mathbf{Q}_{yy}^{-1} \mathbf{b}_{\hat{\mathbf{y}}} = \|\mathbf{b}_{\hat{\mathbf{y}}}\|_{\mathbf{Q}_{yy}}^2. \quad (24.13)$$

Thus under \mathcal{H}_a , the χ^2 -distribution is driven by the influential *bias-to-noise* (BNR) ratio $\lambda_{\hat{\mathbf{y}}}$. Under \mathcal{H}_a , the probability of a hazardous occurrence, i. e., the probability that $\hat{\mathbf{x}}$ lies outside its $100(1 - \eta)\%$ confidence region, is then given as

$$P_H = P\left(\|\hat{\mathbf{x}} - \mathbf{x}\|_{\mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}}^2 \geq \chi_{\eta}^2(n, 0) \mid \mathcal{H}_a\right) \quad (24.14)$$

and this probability will get larger the larger the influential bias-to-noise ratio $\lambda_{\hat{\mathbf{y}}}$ gets.

The hazardous probability P_H depends, next to \mathbf{A} and \mathbf{Q}_{yy} , on \mathbf{C} and \mathbf{b} , i. e., on the type and size of the bias. Hence, (24.14) can now be used to determine the probability of a hazardous occurrence as function of the

bias. In this way one can find out which biases really matter. For instance, if a criterion was set on the maximal allowable P_H , then *inversion* of (24.14) would give the corresponding maximal allowable bias.

Above we have chosen the confidence region as the region against which to judge the quality of the solution. However, instead of the confidence region itself, one can of course take any other \mathbf{x} -centered region R_x . In some cases such a region may be more suitable for the application at hand. Instead of using (24.14), one would then be using $P_H = P(\hat{\mathbf{x}} \notin R_x \mid \mathcal{H}_a)$ to compute the hazardous probability.

We now present a few examples of the influential bias-to-noise ratio.

Example 24.1 Influential outlier

Consider a time series of uncorrelated scalar observations y_i , $i = 1, \dots, k$, all of which are assumed to have the same mean $E(y_i) = x$ and same variance $D(y_i) = \sigma^2$. Then we have the simple linear model with $\mathbf{A} = (1, \dots, 1)^T$ and $\mathbf{Q}_{yy} = \sigma^2 \mathbf{I}_k$. If we consider that an outlier of size b occurred at epoch l , then $\mathbf{b}_y = \mathbf{c}_l b$, with \mathbf{c}_l being the canonical unit vector having the 1 as its l th entry. From this we can compute the variance of the LS estimator as $\sigma_{\hat{\mathbf{x}}}^2 = \sigma^2/k$ and the influential bias as $b_{\hat{\mathbf{x}}} = b/k$. Hence, the influential bias-to-noise ratio (24.13) follows then as

$$\lambda_{\hat{\mathbf{y}}}(\text{outlier}) = \frac{1}{\sqrt{k}} \frac{|b|}{\sigma}. \quad (24.15)$$

This shows that the impact of the outlier on the LS solution becomes less significant as k gets larger.

Example 24.2 Influential slip

In the previous example, the symmetry of the problem was such that the time l of outlier occurrence did not affect its influence on the estimated parameter. Now we assume that a slip of size b occurred at epoch l . Then $\mathbf{b}_y = \mathbf{s}_l b$, with $\mathbf{s}_l = (0, \dots, 0, 1, \dots, 1)^T$ having $(k - l + 1)$ entries equal to 1. From this we can compute the influential bias as $b_{\hat{\mathbf{x}}} = (k - l + 1)b/k$. This shows that the influential bias is largest when $l = 1$ and smallest when $l = k$. In the first case, since \mathbf{A} and \mathbf{s}_l are the same, the bias is nontestable and gets completely passed on to the parameter solution. In the second case, the slip occurs at the last epoch and therefore acts as if it is an outlier. With $\sigma_{\hat{\mathbf{x}}}^2 = \sigma^2/k$, the influential bias-to-noise ratio (24.13) follows as

$$\lambda_{\hat{\mathbf{y}}}(\text{slip}) = \frac{(k - l + 1)}{\sqrt{k}} \frac{|b|}{\sigma}. \quad (24.16)$$

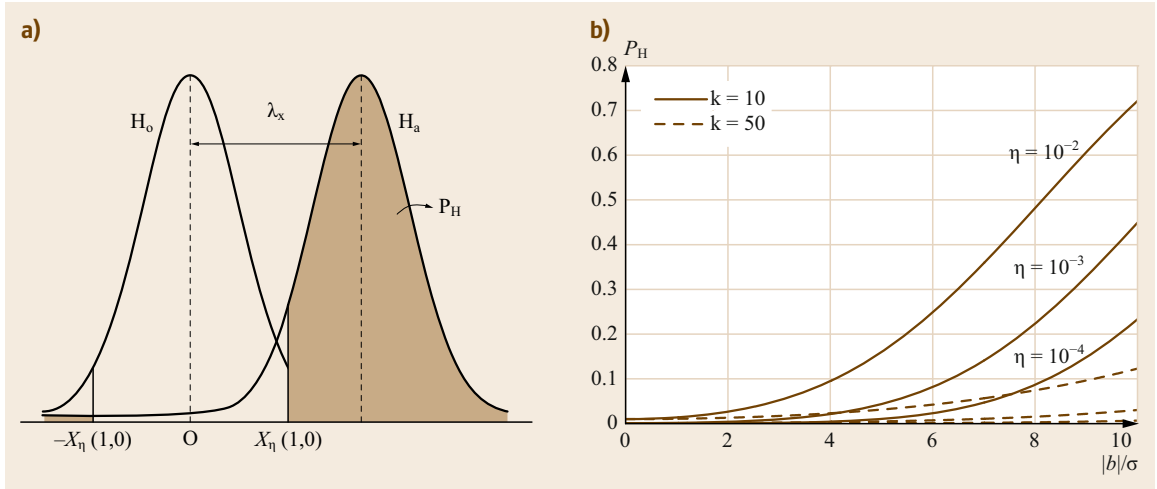


Fig. 24.3 (a) The probability density function (PDF) of $(\hat{x}-x)/\sigma_{\hat{x}}$ under \mathcal{H}_0 and \mathcal{H}_a of Example 24.1. (b) Hazardous probability $P_H = P((\hat{x}-x)^2/\sigma_{\hat{x}}^2 \geq \chi_\eta^2(1,0) \mid \mathcal{H}_a)$ as function of $|b|/\sigma$

This shows, in contrast to (24.15), that the impact of a slip gets more significant as k gets larger.

Example 24.3 Outlier influential BNR

If the variance matrix of the observables is diagonal, i. e., $\mathbf{Q}_{yy} = \text{diag}(\sigma_{y_1}^2, \dots, \sigma_{y_m}^2)$, and matrix \mathbf{C} is the canonical unit vector $\mathbf{c}_i = (0, \dots, 1, \dots, 0)^T$, then the influential BNR (24.13) simplifies to

$$\lambda_{\hat{y}} = \frac{\sigma_{\hat{y}_i} |b|}{\sigma_{y_i} \sigma_{y_i}}. \quad (24.17)$$

Hence, outliers in observations that do not improve much in precision will be most influential.

Example 24.4 P_H and bias

The probability of hazardous occurrence, P_H (24.14), is given in Fig. 24.3 for the case of Example 24.1 as function of $|b|/\sigma$ for different values of η and k . In this

case we have

$$\frac{(\hat{x}-x)}{\sigma_{\hat{x}}} \stackrel{\mathcal{H}_a}{\sim} N(\lambda_{\hat{y}}(\text{outlier}), 1),$$

(24.15) and therefore

$$\begin{aligned} P\left(\frac{(\hat{x}-x)^2}{\sigma_{\hat{x}}^2} \geq \chi_\eta^2(1,0) \mid \mathcal{H}_a\right) \\ = \Phi(-\chi_\eta(1,0) + \lambda_{\hat{y}}) + \Phi(-\chi_\eta(1,0) - \lambda_{\hat{y}}), \end{aligned}$$

with

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}v^2\right) dv.$$

As the figure shows, the probability of hazardous occurrence gets larger for larger biases. It gets smaller if η , the hazardous occurrence under \mathcal{H}_0 is smaller and/or if more epochs of data are used (stronger model).

24.3 Testing for a Bias

24.3.1 The Most Powerful Test Statistic

The aim of testing the working hypothesis \mathcal{H}_0 against the alternative is to obtain protection against hazardous occurrences. Note that with

$$\begin{aligned} E(\mathbf{y}) &= \mathbf{A}\mathbf{x} + \mathbf{C}\mathbf{b}, \\ D(\mathbf{y}) &= \mathbf{Q}_{yy}, \end{aligned}$$

the pair (24.3)–(24.4) can alternatively be written as

$$\mathcal{H}_0 : \mathbf{b} = \mathbf{0} \text{ versus } \mathcal{H}_a : \mathbf{b} \neq \mathbf{0}. \quad (24.18)$$

To decide now whether or not the bias \mathbf{b} can be neglected, it seems reasonable to estimate \mathbf{b} under \mathcal{H}_a and evaluate its significance. The LS solution of \mathbf{b} is given

as

$$\begin{aligned}\hat{\mathbf{b}} &= (\bar{\mathbf{C}}^T \mathbf{Q}_{yy}^{-1} \bar{\mathbf{C}})^{-1} \bar{\mathbf{C}}^T \mathbf{Q}_{yy}^{-1} \mathbf{y}, \\ \bar{\mathbf{C}} &= \mathbf{P}_A^\perp \mathbf{C}.\end{aligned}\quad (24.19)$$

To evaluate its significance, we include the precision of $\hat{\mathbf{b}}$ as well and therefore compare $\hat{\mathbf{b}}$ with its variance matrix $\mathbf{Q}_{\hat{\mathbf{b}}} = (\bar{\mathbf{C}}^T \mathbf{Q}_{yy}^{-1} \bar{\mathbf{C}})^{-1}$. A scalar measure that makes such comparison possible is the (dimensionless) quadratic form

$$T_q = \hat{\mathbf{b}}^T \mathbf{Q}_{\hat{\mathbf{b}}}^{-1} \hat{\mathbf{b}}. \quad (24.20)$$

The estimate $\hat{\mathbf{b}}$ is considered insignificant if the outcome for T_q is small enough. With the estimate $\hat{\mathbf{b}}$ being insignificant, the decision is then made that there is no reason to believe that a bias is present. Thus, \mathcal{H}_0 is accepted (not rejected) if T_q is small enough

$$\text{Accept } \mathcal{H}_0 \text{ if } T_q \leq \chi_\alpha^2(q, 0) \quad (24.21)$$

and rejected otherwise, with $\chi_\alpha^2(q, 0)$ the tolerance (or critical) value that still needs to be chosen. To judge what is large and what is small, we need the distribution of the test statistic T_q . It is given by the noncentral χ^2 -distribution [24.29–31]

$$T_q \sim \chi^2(q, \lambda_\varepsilon^2), \quad (24.22)$$

with noncentrality parameter

$$\begin{aligned}\lambda_\varepsilon^2 &= \mathbf{b}^T \mathbf{Q}_{\hat{\mathbf{b}}}^{-1} \mathbf{b} = \mathbf{b}_\varepsilon^T \mathbf{Q}_{yy}^{-1} \mathbf{b}_\varepsilon \\ &= \|\mathbf{b}_\varepsilon\|_{\mathbf{Q}_{yy}}^2.\end{aligned}\quad (24.23)$$

Note that $\lambda_\varepsilon = 0$ under \mathcal{H}_0 , in which case the test statistic T_q has a central χ^2 -distribution. With the distribution of T_q given, we can compute the probabilities

$$\begin{aligned}P_{\text{FA}} &= P[T_q > \chi_\alpha^2(q, 0) | \mathcal{H}_0] = \alpha, \\ P_{\text{MD}} &= P[T_q \leq \chi_\alpha^2(q, 0) | \mathcal{H}_a] = \beta.\end{aligned}\quad (24.24)$$

The probability of *false alarm*, P_{FA} , is the probability of rejecting the null hypothesis \mathcal{H}_0 while it is true. It is also referred to as the level of significance or size of the test. By choosing a false-alarm rate, say $P_{\text{FA}} = \alpha$, one can compute the corresponding critical value $\chi_\alpha^2(q, 0)$ and execute the test (24.21). The value chosen for α depends on the application. A choice $\alpha = 0.1\%$ for instance, implies that one is willing to accept that the test will have one out of thousand wrongful rejections. The larger α is chosen, the smaller the tolerance value $\chi_\alpha^2(q, 0)$ will be.

When the false alarms are costly, one will tend to choose smaller values for α than otherwise. In classical terrestrial surveying, for instance, false alarms could be costly due to the need of having to do a remeasurement in case observations were rejected. The situation with GNSS is quite different. Observations are relatively cheap to come by and with today's high sampling rates, there is also an abundance of observations. Hence, with many of today's GNSS applications the level of significance can be chosen as larger than what the classical surveyor was used to.

Only *rejecting* or *not rejecting* the null hypothesis is not always informative. One may also want to know the strength of evidence on which the decision is made. Hence, one could ask for every α whether the test rejects at that level. The smallest α at which the test rejects is called the *p-value*. It is defined as the probability under \mathcal{H}_0 of obtaining an outcome of the test statistic equal to or more extreme than what was actually observed,

$$p(T_q) = \int_{T_q}^{\infty} f_{\chi^2(q, 0)}(x) dx \quad (24.25)$$

in which $f_{\chi^2(q, 0)}(x)$ is the probability density function (PDF) of the χ^2 -distribution. The smaller the *p-value*, the larger one considers the evidence against \mathcal{H}_0 .

The probability of *missed detection*, P_{MD} , is the probability of wrongfully accepting the null hypothesis. The complement to the probability of missed detection, $P_{\text{power}} = 1 - P_{\text{MD}} = \gamma$, is known as the *power* of the test.

The missed detection probability P_{MD} depends, through $\chi_\alpha^2(q, 0)$, on the chosen false-alarm rate α , and, through the distribution of T_q under \mathcal{H}_a , it also depends on the testable bias-to-noise ratio λ_ε (24.23). It thus depends on the bias \mathbf{b} and on the precision with which it can be estimated, $\mathbf{Q}_{\hat{\mathbf{b}}}$. P_{MD} is monotone decreasing with λ_ε . Thus the larger the bias and/or the more precise it can be estimated, the smaller the missed detection probability P_{MD} becomes.

Preferably one would like both error probabilities, P_{FA} and P_{MD} , to be small. Unfortunately, however, it is not possible to minimize both simultaneously. If α is chosen smaller, then β will get larger, and vice versa (Fig. 24.4). This fundamental trade-off in hypothesis testing motivated *Neyman and Pearson* [24.32] to develop tests that maximize power $\gamma = 1 - \beta$ for given false alarm α . The above test (24.21), which is a *uniformly most powerful invariant (UMPI)* test, can be shown to be optimal in this sense [24.29, 30]. It thus has, for a fixed false alarm, the smallest missed detection probability of all invariant tests.

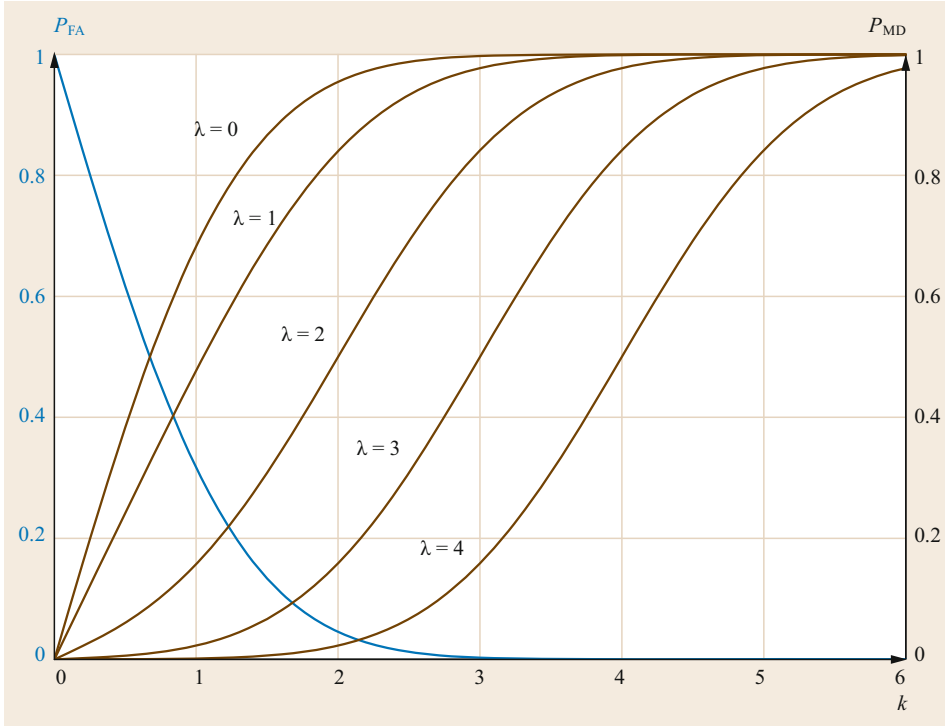


Fig. 24.4 False alarm and missed detection probabilities, $P_{FA} = 2[1 - \Phi(k)]$ and $P_{MD} = [\Phi(k - \lambda) + \Phi(k + \lambda) - 1]$, as function of k and λ , for testing $\mathcal{H}_0 : \lambda = 0$ versus $\mathcal{H}_a : \lambda \neq 0$ with test statistic $T_{q=1} \sim \chi^2(1, \lambda^2)$ and rejection interval $T_{q=1} > k^2$. For a smaller probability of false alarm, $P_{FA} = \alpha$, the probability of missed detection, $P_{MD} = \beta$, gets larger. For a fixed P_{FA} (or k), the missed detection probability gets smaller for larger λ

We now present a few examples on the testable bias-to-noise ratio.

Example 24.5 Testable outlier

For the case of Example 24.1, one can compute the testable bias-to-noise ratio (24.23) for an outlier as

$$\lambda_{\hat{e}}(\text{outlier}) = \left(1 - \frac{1}{k}\right)^{\frac{1}{2}} \frac{|b|}{\sigma}. \quad (24.26)$$

Hence, the outlier is nontestable in the case $k = 1$, which of course corresponds to the case when redundancy is lacking. ■

Example 24.6 Outlier P_{MD}

For the case of Example 24.1, with an assumed outlier in the j th observation, the outlier test statistic reads

$$T_{q=1} = \left(\frac{\hat{b}}{\sigma_{\hat{b}}}\right)^2, \quad (24.27)$$

with

$$\hat{b} = \frac{k}{k-1} (y_j - \bar{y})$$

and variance

$$\sigma_{\hat{b}}^2 = \frac{k}{k-1} \sigma^2$$

where \bar{y} is the average of the k observations. The probability of missed detection P_{MD} for the outlier test $T_{q=1} > \chi_{\alpha}^2(1, 0)$ is shown in Fig. 24.5 for different values of k and α . ■

Example 24.7 Testable slip

For the case of Example 24.2, one can compute the testable bias-to-noise ratio for the slip as

$$\lambda_{\hat{e}}(\text{slip}) = \left[(l-1) \left(1 - \frac{l-1}{k}\right)\right]^{\frac{1}{2}} \frac{|b|}{\sigma}. \quad (24.28)$$

Hence, the slip is nontestable for $l = 1$ and $k = 1$. It is nontestable in the first case, since the slip gets then completely absorbed by the parameters solution, while in the second case it is nontestable due to lack of redundancy. The slip is best testable if it occurs halfway into the time series at the epoch being the nearest integer of

$$\frac{1}{2}k + 1 = \arg \max_l \lambda_{\hat{e}}.$$

■

Example 24.8 Outlier testable BNR

If the variance matrix of the observables is diagonal, i. e., $\mathbf{Q}_{yy} = \text{diag}(\sigma_{y_1}^2, \dots, \sigma_{y_m}^2)$, and matrix \mathbf{C} is the

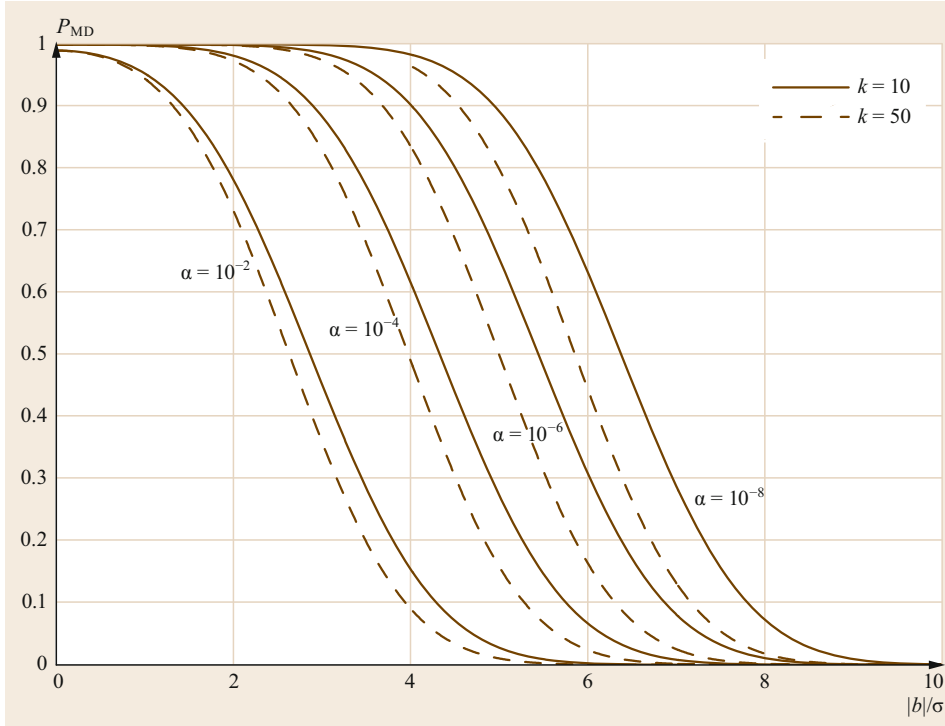


Fig. 24.5 Probability of missed detection P_{MD} , for the outlier test of Examples 24.1, 24.5 and 24.6, as function of $|b|/\sigma$, shown for different number of epochs k (10, 50) and different levels of significance α

canonical unit vector $\mathbf{c}_i = (0, \dots, 1, \dots, 0)^T$, then the testable BNR (24.23) simplifies to

$$\lambda_{\hat{\mathbf{e}}} = \left[1 - \left[\frac{\sigma_{\hat{y}_i}}{\sigma_{y_i}} \right]^2 \right]^{\frac{1}{2}} \frac{|b|}{\sigma_{y_i}}. \quad (24.29)$$

Hence, outliers in observations that do not improve much in precision will be poorly testable. ■

24.3.2 Alternative Expressions for Test Statistic T_q

Alternative expressions exist for the UMPI test statistic (24.21). They provide extra insight and give alternative ways of computing the test statistic. Here we present three such geometric expressions taken from [24.29]. First note that T_q can be interpreted as the squared weighted norm of the estimated bias vector,

$$T_q = \hat{\mathbf{b}}^T \mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{b}}}^{-1} \hat{\mathbf{b}} = \|\hat{\mathbf{b}}\|_{\mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{b}}}}^2. \quad (24.30)$$

This shows that T_q measures the square of the *length* of $\hat{\mathbf{b}}$. Three alternative ways of expressing the test statistic

in geometric terms are (Fig. 24.6)

$$\begin{aligned} \|\hat{\mathbf{b}}\|_{\mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{b}}}}^2 &= \|\mathbf{P}_{\bar{\mathbf{C}}}\mathbf{y}\|_{\mathbf{Q}_{yy}}^2 \\ &= \|\hat{\mathbf{y}} - \hat{\mathbf{y}}_a\|_{\mathbf{Q}_{yy}}^2 \\ &= \|\hat{\mathbf{e}}\|_{\mathbf{Q}_{yy}}^2 - \|\hat{\mathbf{e}}_a\|_{\mathbf{Q}_{yy}}^2. \end{aligned} \quad (24.31)$$

The first expression follows from substitution of

$$\hat{\mathbf{b}} = (\bar{\mathbf{C}}^T \mathbf{Q}_{yy}^{-1} \bar{\mathbf{C}})^{-1} \bar{\mathbf{C}}^T \mathbf{Q}_{yy}^{-1} \mathbf{y},$$

with $\bar{\mathbf{C}} = \mathbf{P}_A^\perp \mathbf{C}$, (24.19), into (24.30). The second expression follows from the first by using the projector decomposition $\mathbf{P}_{[A,C]} = \mathbf{P}_A + \mathbf{P}_{\bar{\mathbf{C}}}$, thereby recognizing that $\hat{\mathbf{y}} = \mathbf{P}_A \mathbf{y}$ and $\hat{\mathbf{y}}_a = \mathbf{P}_{[A,C]} \mathbf{y}$ are the LS estimators of the mean of \mathbf{y} under \mathcal{H}_0 and \mathcal{H}_a respectively. This second expression of (24.31) shows that the test statistic is the squared norm of the *solution separation* between the two hypotheses. Finally, the third expression of (24.31) follows from the second by using

$$\mathbf{P}_A^\perp = \mathbf{P}_{[A,C]}^\perp + \mathbf{P}_{\bar{\mathbf{C}}},$$

thereby recognizing that

$$\begin{aligned} \hat{\mathbf{e}} &= \mathbf{P}_A^\perp \mathbf{y}, \\ \hat{\mathbf{e}}_a &= \mathbf{P}_{[A,C]}^\perp \mathbf{y}, \end{aligned}$$

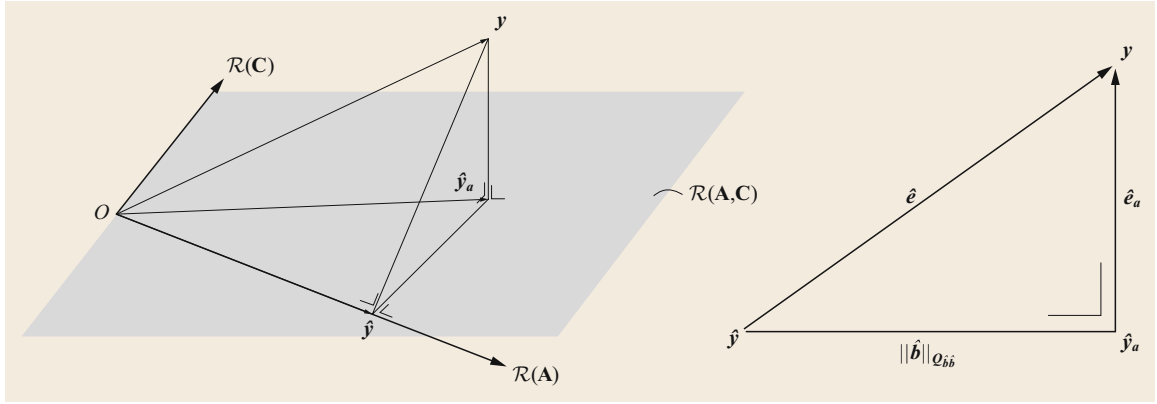


Fig. 24.6 The geometry of the UMPI test statistic $T_q = \|\hat{\mathbf{b}}\|_{\mathbf{Q}_{bb}}^2$ (after [24.29])

and

$$\mathcal{R}(\mathbf{A}, \mathbf{C})^\perp \perp \mathcal{R}(\bar{\mathbf{C}}).$$

This third expression shows that the test statistic can also be written as the difference of the squared norm LS residuals under \mathcal{H}_0 and \mathcal{H}_a respectively. Hence,

$$T_q = \|\hat{\mathbf{b}}\|_{\mathbf{Q}_{bb}}^2$$

is the value by which the weighted sum-of-squared residuals gets reduced when going from \mathcal{H}_0 to \mathcal{H}_a .

24.3.3 Test Statistic T_q Expressed in LS Residuals

When (24.30), or one of the last two expressions of (24.31), is used for the computation of T_q , an explicit estimation under \mathcal{H}_a is needed to obtain $\hat{\mathbf{b}}$, $\hat{\mathbf{y}}_a$ or $\hat{\mathbf{e}}_a$ respectively. This can be avoided if one makes use of the first expression of (24.31), thereby recognizing that $\mathbf{P}_{\bar{\mathbf{C}}}\mathbf{y} = \mathbf{P}_{\bar{\mathbf{C}}}\hat{\mathbf{e}}$. We have

$$\begin{aligned} T_q &= \|\mathbf{P}_{\bar{\mathbf{C}}}\hat{\mathbf{e}}\|_{\mathbf{Q}_{yy}}^2 \\ &= \hat{\mathbf{e}}^T \mathbf{Q}_{yy}^{-1} \mathbf{C} (\mathbf{C}^T \mathbf{Q}_{yy}^{-1} \mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}} \mathbf{Q}_{yy}^{-1} \mathbf{C})^{-1} \mathbf{C}^T \mathbf{Q}_{yy}^{-1} \hat{\mathbf{e}}. \end{aligned} \quad (24.32)$$

This is the expression most often used for computing the test statistic, since $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$ and $\mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}} = \mathbf{Q}_{yy} - \mathbf{Q}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ are usually readily available when computing under \mathcal{H}_0 . Thus only the $m \times q$ matrix \mathbf{C} in (24.32) changes when one changes to a different \mathcal{H}_a . We now consider the two special cases $q = m - n$ and $q = 1$.

The $q = m - n$ Case

In this case the matrix $[\mathbf{A}, \mathbf{C}]$ of \mathcal{H}_a (24.4) is square and invertible, implying that its range space coincides with

the complete observation space, $\mathcal{R}[\mathbf{A}, \mathbf{C}] = \mathbb{R}^m$. Hence, no restrictions are then put on the mean $E(\mathbf{y})$ of \mathbf{y} under \mathcal{H}_a . As a consequence, the least-squares residual vector under \mathcal{H}_a will be zero, $\hat{\mathbf{e}}_a = \mathbf{0}$, from which it follows, using the last expression in (24.31), that the UMPI test statistic simplifies to

$$T_{q=m-n} = \|\hat{\mathbf{e}}\|_{\mathbf{Q}_{yy}}^2. \quad (24.33)$$

This statistic is used when one wants to test the null hypothesis against the most relaxed hypothesis $E(\mathbf{y}) \in \mathbb{R}^m$. Its expression can be further simplified in the case of a diagonal variance matrix of the observables. We then have $T_{q=m-n} = \sum_{i=1}^m (\hat{e}_i / \sigma_{y_i})^2$, which is often referred to as the (weighted) residual sum of squares (RSSs).

The $q = 1$ Case

In the one-dimensional case, matrix \mathbf{C} becomes a vector \mathbf{c} and the test statistic can be written as the square of a normally distributed variable w ,

$$T_{q=1} = (w)^2, \quad (24.34)$$

with

$$w = \frac{\mathbf{c}^T \mathbf{Q}_{yy}^{-1} \hat{\mathbf{e}}}{\sqrt{\mathbf{c}^T \mathbf{Q}_{yy}^{-1} \mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}} \mathbf{Q}_{yy}^{-1} \mathbf{c}}} \quad (24.35)$$

distributed as

$$w \stackrel{\mathcal{H}_0}{\sim} \mathcal{N}(0, 1); w \stackrel{\mathcal{H}_a}{\sim} \mathcal{N}(\lambda_{\hat{\mathbf{e}}, q=1}, 1).$$

Hence, for $q = 1$, test (24.21) can be equivalently written as

$$\text{Accept } \mathcal{H}_0 \text{ if } |w| \leq \mathcal{N}_{\alpha/2}(1, 0). \quad (24.36)$$

This statistic (24.35) is known as Baarda's w -test statistic [24.33, 34]. It is used when one has a specific one-dimensional bias in mind. The specification is then made through the choice of the \mathbf{c} -vector. The acceptance region of (24.36) is an interval. The acceptance interval becomes a region in the case where more than one w -test is used, each with its own \mathbf{c} -vector (Fig. 24.7).

The expression for the w -statistic simplifies considerably in the case where the observables are uncorrelated (i. e., $\mathbf{Q}_{yy} = \text{diag}(\sigma_{y_1}^2, \dots, \sigma_{y_m}^2)$) and the \mathbf{c} -vector is of canonical unit-vector form as used for outlier identification (i. e., $\mathbf{c} = (0, \dots, 0, 1, 0, \dots, 0)^T$, with the 1 at its i th entry). Then, since

$$\mathbf{c}^T \mathbf{Q}_{yy}^{-1} \hat{\mathbf{e}} = \frac{\hat{e}_i}{\sigma_{y_i}}$$

and

$$(\mathbf{c}^T \mathbf{Q}_{yy}^{-1} \mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}} \mathbf{Q}_{yy}^{-1} \mathbf{c})^{\frac{1}{2}} = \frac{\sigma_{\hat{e}_i}}{\sigma_{y_i}},$$

the general expression (24.35) reduces to the normalized LS residual

$$w = \frac{\hat{e}_i}{\sigma_{\hat{e}_i}}. \quad (24.37)$$

In this case it is simply the LS residual divided by its standard deviation.

24.3.4 Optimality of the w -Test Statistic

We already mentioned that the T_q statistic and therefore for $q = 1$ the w statistic, is optimal in the sense of having the smallest missed detection probability for a given level of significance. This optimality can also be looked at from a geometric point of view. Let

$$v = \frac{\mathbf{f}^T \hat{\mathbf{e}}}{\sqrt{\mathbf{f}^T \mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}} \mathbf{f}}} \quad (24.38)$$

be any normalized linear function of the least-squares residual vector $\hat{\mathbf{e}}$. Then

$$v \stackrel{\mathcal{H}_0}{\sim} \mathcal{N}(0, 1) \text{ and } v \stackrel{\mathcal{H}_a}{\sim} \mathcal{N}\left(\frac{\mathbf{f}^T \mathbf{b}_{\hat{\mathbf{e}}}}{\sqrt{\mathbf{f}^T \mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}} \mathbf{f}}}, 1\right). \quad (24.39)$$

The bias, i. e., nonzero mean, of v under \mathcal{H}_a becomes better detectable the larger it is. Hence, the optimal test statistic v is the one for which \mathbf{f} is chosen to maximize this bias. The solution is

$$\hat{\mathbf{f}} = \mathbf{Q}_{yy}^{-1} \mathbf{P}_A^\perp \mathbf{b}_{\hat{\mathbf{e}}} = \arg \max_{\mathbf{f}} \frac{|\mathbf{f}^T \mathbf{b}_{\hat{\mathbf{e}}}|}{\sqrt{\mathbf{f}^T \mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}} \mathbf{f}}}, \quad (24.40)$$

which indeed shows that $w = \hat{\mathbf{f}}^T \hat{\mathbf{e}} / (\hat{\mathbf{f}}^T \mathbf{Q}_{\hat{\mathbf{e}}\hat{\mathbf{e}}} \hat{\mathbf{f}})^{\frac{1}{2}}$.

24.3.5 The Minimal Detectable Bias

The performance of the UMPI test (24.21) is described by its probabilities of false alarm P_{FA} and missed detection P_{MD} respectively. The probability of missed

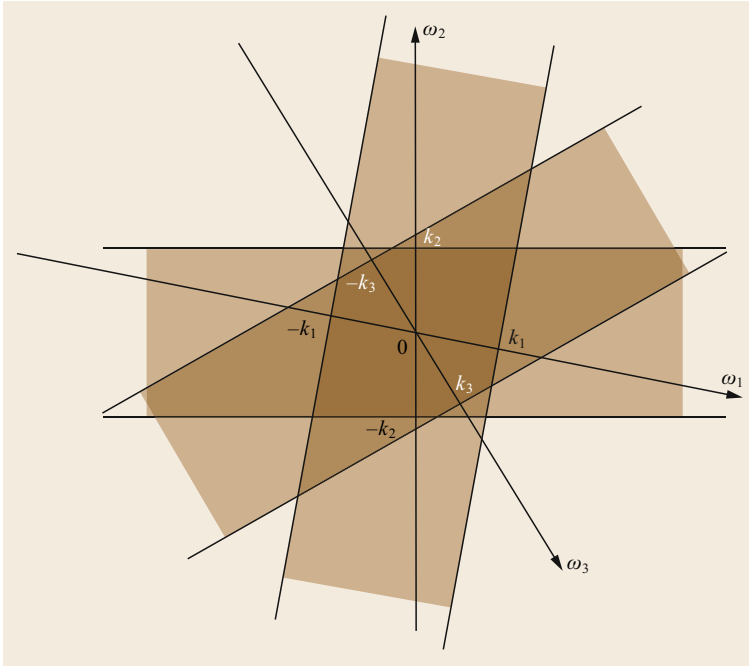


Fig. 24.7 Intersection of multiple w -test acceptance regions (after [24.33])

detection P_{MD} (24.24) can be computed for any type and size of bias. One can however also follow the reverse route and determine the bias that can at least be found for a certain probability of missed detection, say $P_{MD} = \beta$. The steps are then as follows. From setting α , β , and q , one computes through *inversion* the corresponding testable bias-to-noise ratio, denoted as $\lambda_{\hat{e}}(\alpha, \beta, q)$, the value of which provides then a yardstick for the testable bias,

$$\lambda_{\hat{e}}^2(\alpha, \beta, q) = \|\mathbf{b}_{\hat{e}}\|_{\mathbf{Q}_{yy}}^2 = \|\mathbf{P}_A^\perp \mathbf{C} \mathbf{b}\|_{\mathbf{Q}_{yy}}^2. \quad (24.41)$$

This quadratic equation describes a q -dimensional ellipsoid for the bias \mathbf{b} . Biases outside this region will have a larger than $\gamma = 1 - \beta$ probability of being detected under \mathcal{H}_a , while biases inside it will have a smaller probability. A further inversion is possible to find the corresponding biases themselves. Let \mathbf{u} be a unit vector ($\|\mathbf{u}\| = 1$) and parametrize the bias vector as $\mathbf{b} = \|\mathbf{b}\| \mathbf{u}$. Substitution into (24.41), followed by inversion gives then

$$\mathbf{b} = \sqrt{\frac{\lambda_{\hat{e}}^2(\alpha, \beta, q)}{\|\mathbf{P}_A^\perp \mathbf{C} \mathbf{u}\|_{\mathbf{Q}_{yy}}^2}} \mathbf{u} \text{ with } \|\mathbf{u}\| = 1. \quad (24.42)$$

This is the vectorial form of Baarda's *minimal detectable bias* (MDB) [24.33, 34]. Baarda referred to his MDBs as *boundary values* (in Dutch: grenswaarden); the nowadays more customary term MDB was introduced in [24.35].

The length $\|\mathbf{b}\|$ of the MDB vector is the smallest size of bias vector that can be found with probability $\gamma = 1 - \beta$ in the direction \mathbf{u} using the UMPI test (24.21). By letting \mathbf{u} vary over the unit sphere in \mathbb{R}^q one obtains the whole range of MDBs that can be detected with probability γ . Baarda in his work on the strength analysis of general purpose networks, applied his general MDB-form to data snooping [24.36]. Applications of the vectorial form can be found, for example, in [24.37, 38] for deformation analysis and in [24.39, 40] for trend testing. MDB analyses for the Global Positioning System (GPS) can be found in [24.41–43], for comparing and combining GNSSs in [24.44–52], and for recursive testing with applications to navigation in [24.35, 53].

Outlier MDB

For outliers ($q = 1$), the MDB expression simplifies when the variance matrix is diagonal,

$$\mathbf{Q}_{yy} = \text{diag}(\sigma_{y_1}^2, \dots, \sigma_{y_m}^2).$$

The MDB of the i th observable reads then

$$\|\mathbf{b}_i\| = \sigma_{y_i} \sqrt{\frac{\lambda_{\hat{e}}^2(\alpha, \beta, q = 1)}{1 - \frac{\sigma_{y_i}^2}{\sigma_{y_i}^2}}}, \quad i = 1, \dots, m. \quad (24.43)$$

This shows that the outlier MDB is small if $\sigma_{y_i}^2$ is small compared to $\sigma_{y_i}^2$. On the other hand, it will be large if the precision of the adjusted observation has not improved by much, i.e., if $\sigma_{y_i}^2$ is close to $\sigma_{y_i}^2$. The dimensionless number

$$r_i = 1 - \frac{\sigma_{y_i}^2}{\sigma_{y_i}^2}$$

is called the *local redundancy number*. Since $\sigma_{y_i}^2 \leq \sigma_{y_i}^2$, we have $0 \leq r_i \leq 1$. The reason why r_i is called the local redundancy number is due to the property that they add up to the redundancy itself

$$\sum_{i=1}^m r_i = m - n.$$

This can be shown by using the projector property of \mathbf{P}_A [24.29]. If we replace r_i by its average

$$\bar{r} = \frac{m - n}{m},$$

we get the rough MDB approximation

$$\|\mathbf{b}_i\| \approx \sigma_{y_i} \sqrt{\frac{\lambda_{\hat{e}}^2(\alpha, \beta, q = 1)}{\frac{(m-n)}{m}}}, \quad i = 1, \dots, m. \quad (24.44)$$

Example 24.9 Outlier MDB

Consider the null hypothesis \mathcal{H}_0 : $E(\mathbf{y}) = \mathbf{A}\mathbf{x}$, $\mathbf{Q}_{yy} = \sigma^2 \mathbf{I}_m$, with $m \times 2$ design matrix

$$\mathbf{A} = \begin{bmatrix} 1 & a_1 \\ \vdots & \vdots \\ 1 & a_m \end{bmatrix}. \quad (24.45)$$

To determine the outlier MDB, we first need

$$\sigma_{y_i}^2 = \mathbf{c}_i^T \mathbf{A} \mathbf{Q}_{\hat{\mathbf{x}}} \mathbf{A}^T \mathbf{c}_i$$

(cf. (24.43)). With the variance matrix of the LS solution $\hat{\mathbf{x}}$ given as

$$\mathbf{Q}_{\hat{\mathbf{x}}} = \frac{\sigma^2}{\sum_{i=1}^m \bar{a}_i^2} \begin{bmatrix} \frac{1}{m} \sum_{i=1}^m a_i^2 & -a_c \\ -a_c & 1 \end{bmatrix}, \quad (24.46)$$

where

$$\bar{a}_i = a_i - a_c, \quad a_c = \frac{1}{m} \sum_{i=1}^m a_i,$$

the outlier MDB follows then as

$$\|b_i\| = \sigma \sqrt{\frac{\lambda_e^2(\alpha, \beta, q=1)}{1 - \left(\frac{1}{m} + \frac{\bar{a}_i^2}{\sum_{j=1}^m \bar{a}_j^2}\right)}}, \quad i = 1, \dots, m. \quad (24.47)$$

This shows that the MDB gets smaller the smaller \bar{a}_i is. Hence, an outlier in the i th observable is better detectable if its coefficient a_i is closer to the coefficient's mean value a_c [24.29]. ■

MDB and GNSS Geometry

The previous example applies to GNSS in the case where two of the three position coordinates would be constrained. The first column of (24.45) would then correspond to the receiver clock, while the second would refer to the remaining unknown position coordinate. In case all three coordinates are unknown, the $m \times 4$ GNSS design matrix reads

$$\mathbf{A} = \begin{bmatrix} 1 & \mathbf{u}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{u}_m^T \end{bmatrix}, \quad (24.48)$$

with \mathbf{u}_i being the unit direction vector from receiver to satellite i . The GNSS code-outlier MDB follows then as [24.54]

$$\|b_i\| = \sigma \sqrt{\frac{\lambda_e^2(\alpha, \beta, q=1)}{1 - \left[\frac{1}{m} + \mathbf{c}_i^T \mathbf{P}_{\bar{\mathbf{G}}} \mathbf{c}_i\right]}}, \quad i = 1, \dots, m, \quad (24.49)$$

with projector

$$\mathbf{P}_{\bar{\mathbf{G}}} = \bar{\mathbf{G}}(\bar{\mathbf{G}}^T \bar{\mathbf{G}})^{-1} \bar{\mathbf{G}}^T,$$

centered geometry matrix

$$\begin{aligned} \bar{\mathbf{G}} &= \mathbf{G} - \mathbf{e} \mathbf{g}_c^T, \\ \mathbf{G} &= (\mathbf{u}_1, \dots, \mathbf{u}_m)^T, \\ \mathbf{g}_c^T &= \frac{1}{m} \mathbf{e}^T \mathbf{G}, \end{aligned}$$

and $\mathbf{e} = (1, \dots, 1)^T$ the vector of ones. Compare (24.49) with (24.47).

We can now infer, using (24.49), what the impact of the receiver-satellite geometry is on the detectability of code outliers. A good geometry is one where

$$\mathbf{c}_i^T \mathbf{P}_{\bar{\mathbf{G}}} \mathbf{c}_i = 0.$$

As

$$\mathbf{c}_i^T \mathbf{P}_{\bar{\mathbf{G}}} \mathbf{c}_i = \|\mathbf{u}_i - \mathbf{g}_c\|_{\bar{\mathbf{G}}^T \bar{\mathbf{G}}}^2,$$

it follows the closer \mathbf{u}_i is to the average \mathbf{g}_c in the metric of $(\bar{\mathbf{G}}^T \bar{\mathbf{G}})^{-1}$, the smaller the MDB.

A poor geometry for detecting an outlier in the i th code observable is one where the unit direction vectors \mathbf{u}_j of all other code observables, $j \neq i$, lie on a cone. In that case we have for all

$$j \neq i: \mathbf{u}_j^T \mathbf{a} = 1$$

for some vector \mathbf{a} (the symmetry axis of the cone). Therefore, \mathbf{c}_i would lie in the range space of \mathbf{A} , $\mathbf{c}_i \in \mathcal{R}(\mathbf{A})$, and thus

$$\mathbf{c}_i^T \mathbf{P}_{\mathbf{A}} \mathbf{c}_i = 1,$$

where $\mathbf{A} = [\mathbf{e}, \mathbf{G}]$. The MDB $\|b_i\|$ would then be infinite.

The following is an actual example that demonstrates the above.

Example 24.10 GNSS code-outlier MDBs

This example shows how the receiver-satellite geometry affects the MDBs. Figure 24.8 shows the distribution of six GPS satellites and Table 24.1 shows the MDBs of a single-epoch, code-only solution using these satellites. In this case the average horizontal precision in terms of the horizontal dilution of precision was good, HDOP ≈ 1.23 . The results of Table 24.1 show however that a sufficient precision in the position solution need not necessarily correspond with a sufficiently small MDB. When the skyplot of Fig. 24.8 is compared with Table 24.1, the following remarks can be made. First note that satellites pseudorandom noise (PRN) code 6 and PRN 16 are close together. The two pseudoranges to these two satellites therefore check each other well. This explains the relatively small and almost equal

Table 24.1 Code-outlier MDBs of six GPS satellites having the receiver-satellite geometry of Fig. 24.8

PRN	$ b /\sigma_p$
16	5.41
18	8.81
2	5.25
9	69.55
6	5.62
17	22.63

MDBs for the first and fifth pseudorange. Also note that in the absence of satellite PRN 2, one would expect, because of symmetry, the MDBs of the pseudoranges to PRNs 18 and 17 to be of the same order. With satellite PRN 2 however, additional redundancy for checking pseudorange to PRN 18 enters, which explains why its MDB is smaller than that of the pseudorange to PRN 17. Finally note the large value for the MDB of PRN 9. This is due to the fact that all unit direction vectors, except that of PRN 9, approximately lie on a common cone, the symmetry axis of which is indicated with a *star* in the skyplot of Fig. 24.8. ■

Example 24.11 MDB of nonoptimal test

The derivation and computation of the MDB is not restricted to optimal tests. It can be done for any test and thus for the nonoptimal v test statistic of (24.38) as well. Using a similar derivation as before, its MDB follows as

$$|b|_v = \sqrt{\frac{\lambda_{\hat{e}}^2(\alpha, \beta, 1)}{\|\mathbf{P}_A^\perp \mathbf{c}\|_{\mathbf{Q}_{yy}}^2 \cos^2 \theta}}, \quad (24.50)$$

with the angle defined as $\theta = \angle(\mathbf{Q}_{yy} \mathbf{f}, \mathbf{P}_A^\perp \mathbf{c})$. This indeed shows that the MDB is smallest if \mathbf{f} is chosen as

$\hat{\mathbf{f}}$ of (24.40), in which case it becomes identical to the MDB of the w -test. ■

24.3.6 Hazardous Missed Detection

False alarm (FA) and missed detection (MD) are the two possible incorrect outcomes of the test (24.21). To be able to evaluate their significance, we take the occurrences of hazardous solutions into account as well. Consider therefore Fig. 24.9. It has along the horizontal axis the parameter statistic

$$P_n = \|\hat{\mathbf{x}} - \mathbf{x}\|_{\mathbf{Q}_{\hat{\mathbf{x}}}}^2$$

and along the vertical axis the test statistic

$$T_q = \|\hat{\mathbf{b}}\|_{\mathbf{Q}_{\hat{\mathbf{b}}}}^2.$$

These two random variables are independent, since $\hat{\mathbf{x}}$ and $\hat{\mathbf{e}}$ are independent. With the use of $\chi_{\eta}^2(n, 0)$ and $\chi_{\alpha}^2(q, 0)$, the sample space of the two statistics is divided into four rectangular regions. Figure 24.9a shows the situation under \mathcal{H}_0 and Fig. 24.9b shows the situation under \mathcal{H}_a .

Under \mathcal{H}_0 , there is a false alarm region (FA) and a hazardous region (H) (Fig. 24.9a). The hazardous part

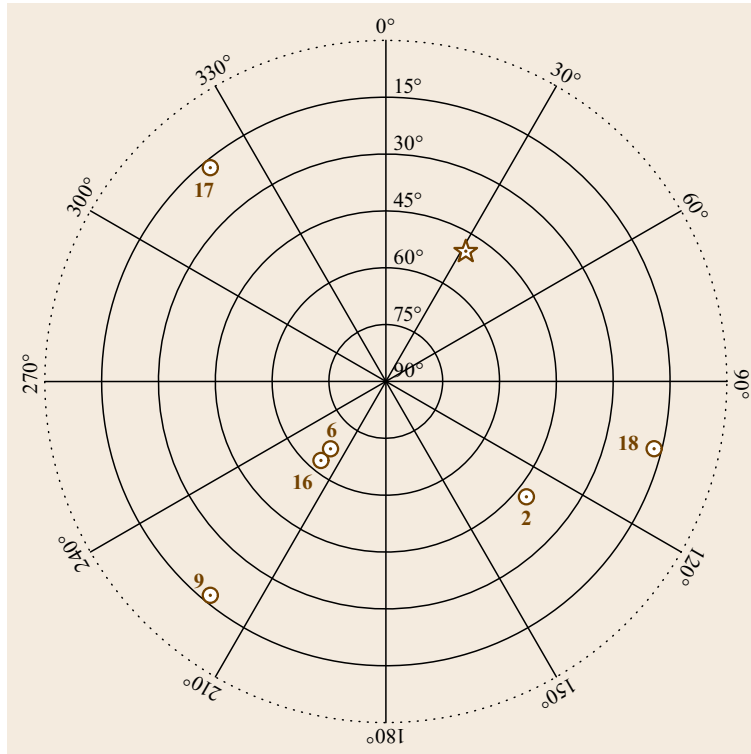


Fig. 24.8 GPS skyplot of Example 10's PRNs 2, 6, 9, 16, 17, 18. The *star* indicates the symmetry axis of the cone determined by all PRNs excluding PRN 9

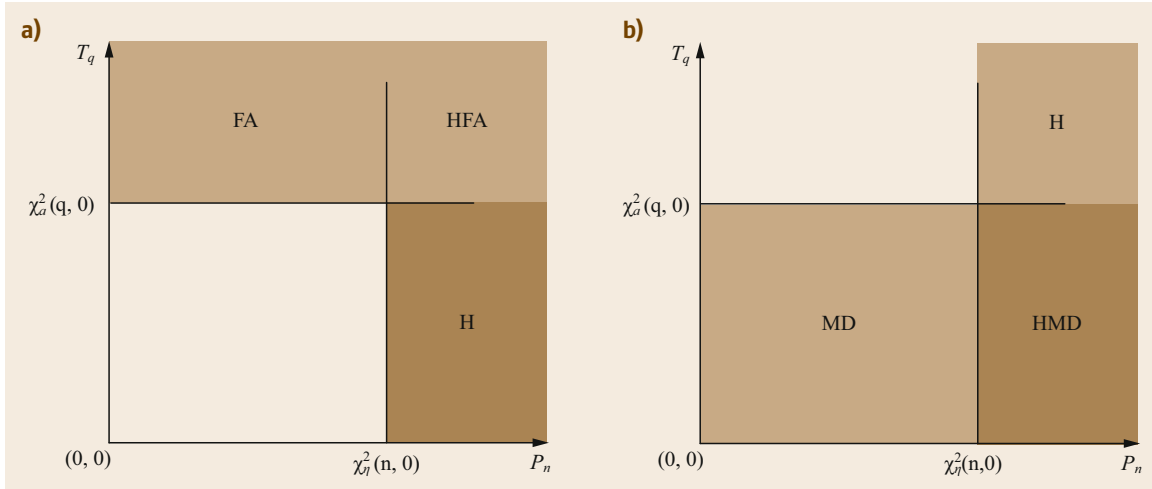


Fig. 24.9a,b The joint (P_n, T_q) -sample space of the parameter statistic $P_n = \|\hat{\mathbf{x}} - \mathbf{x}\|_{\mathbf{Q}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}}^2$ and the test statistic $T_q = \|\hat{\mathbf{b}}\|_{\mathbf{Q}_{\hat{\mathbf{b}}\hat{\mathbf{b}}}}^2$. The sample space has been divided into four regions both under \mathcal{H}_0 (a) and \mathcal{H}_a (b) (H = hazardous, FA = false alarm, MD = missed detection)

of FA should not be too much of a concern, since even though the null hypothesis is falsely rejected, outcomes in the hazardous region is what one would like to avoid anyway. The probability of occurrence of the remaining hazardous part is $P_H \times (1 - P_{FA}) = [\eta \times (1 - \alpha)]\%$. This probability is controlled by choosing η small enough. The remaining region one should be concerned about under \mathcal{H}_0 is the nonhazardous FA region. This is the region for which one rejects the working hypothesis, while having solutions that are nonhazardous and therefore acceptable. The probability of this happening is given by the product $(1 - P_H) \times P_{FA} = [(1 - \eta) \times \alpha]100\%$. This probability is usually controlled by choosing α sufficiently small.

Under \mathcal{H}_a , also the missed detection (MD) region can be further divided into a hazardous and nonhazardous part (Fig. 24.9b). Outcomes in the nonhazardous part are not a problem, because it would mean that despite the presence of undetected biases, their impact is such that the parameter solutions are still acceptable. What is unacceptable however, are missed detections that result in hazardous solutions. The probability that this happens is given by the product

$$P_{HMD} = P_H \times P_{MD}. \quad (24.51)$$

Recall that the hazardous probability P_H is driven by the influential BNR $\lambda_{\hat{\mathbf{y}}}$ (24.13), while the missed detection probability P_{MD} is driven by the testable BNR $\lambda_{\hat{\mathbf{e}}}$ (24.23), both of which in their turn are driven by the actual BNR $\lambda_{\mathbf{y}} = \|\mathbf{b}_{\mathbf{y}}\|_{\mathbf{Q}_{\mathbf{y}\mathbf{y}}}$. The BNR relation is given by the Pythagorean decomposition (Fig. 24.10)

$$\lambda_{\mathbf{y}}^2 = \lambda_{\hat{\mathbf{y}}}^2 + \lambda_{\hat{\mathbf{e}}}^2, \quad (24.52)$$

with actual, influential and testable BNRs,

$$\begin{aligned} \lambda_{\mathbf{y}} &= \|\mathbf{b}_{\mathbf{y}}\|_{\mathbf{Q}_{\mathbf{y}\mathbf{y}}}, \\ \lambda_{\hat{\mathbf{y}}} &= \|\mathbf{b}_{\hat{\mathbf{y}}}\|_{\mathbf{Q}_{\mathbf{y}\mathbf{y}}} = \|\mathbf{b}_{\mathbf{y}}\|_{\mathbf{Q}_{\mathbf{y}\mathbf{y}}} \cos(\phi), \\ \lambda_{\hat{\mathbf{e}}} &= \|\mathbf{b}_{\hat{\mathbf{e}}}\|_{\mathbf{Q}_{\mathbf{y}\mathbf{y}}} = \|\mathbf{b}_{\mathbf{y}}\|_{\mathbf{Q}_{\mathbf{y}\mathbf{y}}} \sin(\phi). \end{aligned} \quad (24.53)$$

The BNRs $\lambda_{\hat{\mathbf{e}}}$ and $\lambda_{\hat{\mathbf{y}}}$ were introduced by Baarda as his measures of *internal* and *external* reliability, respectively [24.33, 34, 36], see also [24.55–59].

As the angle ϕ in (24.53) determines how much of the actual bias is testable and influential respectively, it determines the ratio of the testable BNR to the influential BNR: $\lambda_{\hat{\mathbf{e}}} = \lambda_{\hat{\mathbf{y}}} \tan(\phi)$. The smaller the angle ϕ , the more $\mathbf{b}_{\mathbf{y}}$ and $\mathcal{R}(\mathbf{A})$ are aligned and the more influential the bias will be. The angle itself is determined by the type of bias (i. e., matrix \mathbf{C}) and the strength of the underlying model (i. e., matrices \mathbf{A} and $\mathbf{Q}_{\mathbf{y}\mathbf{y}}$). If we parametrize the bias as before, $\mathbf{b} = \|\mathbf{b}\|\mathbf{u}$, with $\|\mathbf{u}\| = 1$ (cf. (24.42)), then

$$\lambda_{\hat{\mathbf{y}}} = \frac{\|\mathbf{P}_{\mathbf{A}}\mathbf{C}\mathbf{u}\|_{\mathbf{Q}_{\mathbf{y}\mathbf{y}}}}{\|\mathbf{P}_{\mathbf{A}}^{\perp}\mathbf{C}\mathbf{u}\|_{\mathbf{Q}_{\mathbf{y}\mathbf{y}}}} \lambda_{\hat{\mathbf{e}}}. \quad (24.54)$$

To compute and evaluate the probability of hazardous missed detection P_{HMD} one can follow different routes. The first approach goes as follows. From choosing α , β and knowing q , one can first compute the yardstick $\lambda_{\hat{\mathbf{e}}}(\alpha, \beta, q)$. This, together with knowledge of matrix \mathbf{C} (i. e., type of bias), enables the computation of the MDB (24.42), the corresponding BNR $\lambda_{\mathbf{y}}$, and via Pythagoras, the corresponding influential BNR

$$\lambda_{\hat{\mathbf{y}}} = (\lambda_{\mathbf{y}}^2 - \lambda_{\hat{\mathbf{e}}}^2)^{\frac{1}{2}}.$$

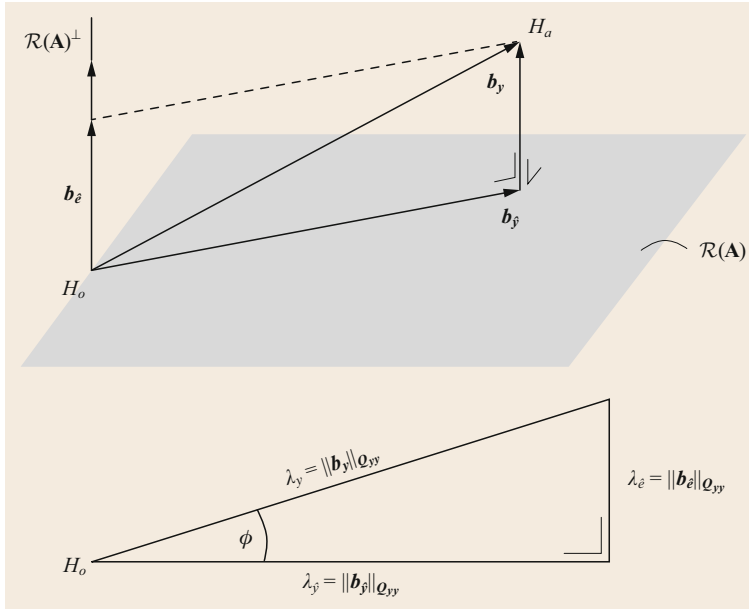


Fig. 24.10 The Pythagorean BNR decomposition $\lambda_y^2 = \lambda_{\hat{y}}^2 + \lambda_{\hat{e}}^2$. The distance between the two hypotheses, \mathcal{H}_0 and \mathcal{H}_a , is given by the actual BNR $\lambda_y = \|E(y|\mathcal{H}_a) - E(y|\mathcal{H}_0)\|_{Q_{yy}} = \|b_y\|_{Q_{yy}}$. Its orthogonal projection onto $\mathcal{R}(\mathbf{A})$ and $\mathcal{R}(\mathbf{A})^\perp$ gives the influential and testable BNRs $\lambda_{\hat{y}}$ and $\lambda_{\hat{e}}$ respectively

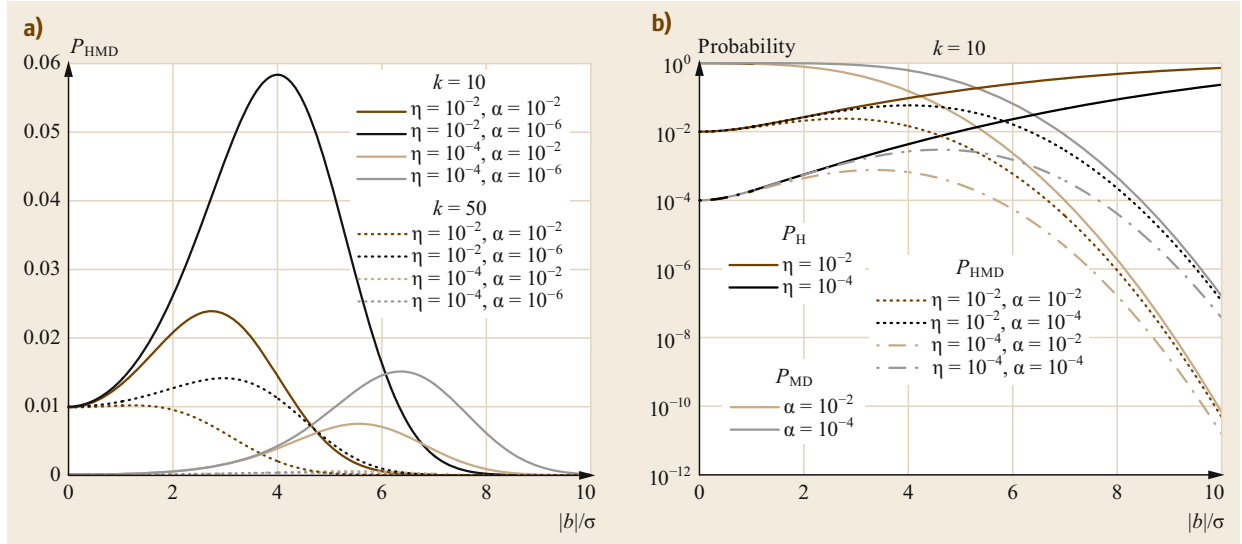


Fig. 24.11a,b Hazardous missed detection probability P_{HMD} as a function of $|b|/\sigma$ for the outlier example (1 and 5) with different settings of k , η , and α (a). Separate graphs of the three probabilities P_H , P_{MD} and P_{HMD} (b)

From it one can then compute the hazardous probability P_H and finally P_{HMD} . This is in essence the procedure as devised by Baarda in [24.34] for designing *general purpose* geodetic networks and for describing their (internal and external) reliability. Thus

$$P_{HMD} = P_H \times \beta$$

in which (24.54) is used for computing P_H , with $\lambda_{\hat{e}} = \lambda_{\hat{e}}(\alpha, \beta, q)$ as a yardstick. Instead of starting with P_{MD}

(or $\lambda_{\hat{e}}$) one can alternatively start with the probability of a hazardous occurrence P_H (or $\lambda_{\hat{y}}$) would such criterion be available, say $P_H = \eta_a$. Then $P_{HMD} = \eta_a \times P_{MD}$ in which the inverse of (24.54) is used for computing P_{MD} from $\lambda_{\hat{y}}$, α and q .

An alternative approach to the above two would be to directly evaluate the probability of hazardous missed detection as function of the bias, $P_{HMD}(b)$, in which case explicit choices of $P_{MD} = \beta$ or $P_H = \eta_a$ are not involved. As P_{MD} gets smaller, but P_H larger, for larger

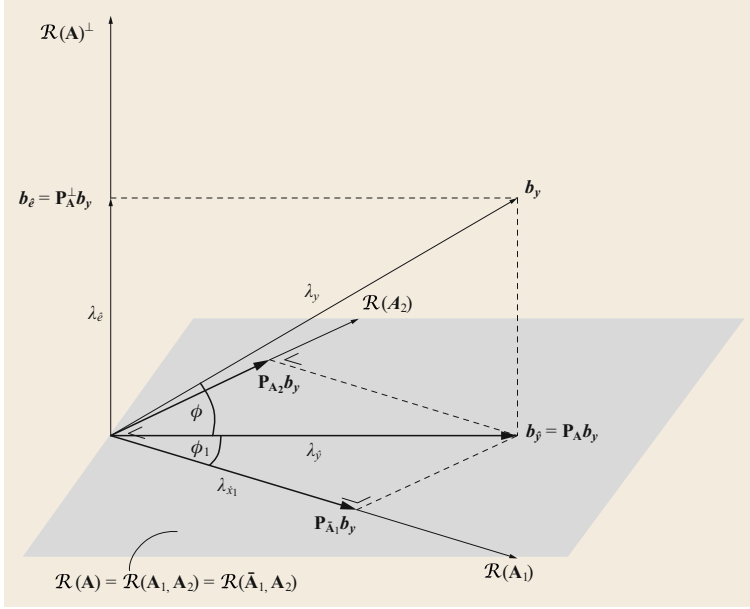


Fig. 24.12 Pythagorean BNR decomposition $\lambda_{\hat{y}}^2 = \lambda_{\hat{x}_1}^2 + \lambda_{\hat{x}_2|x_1}^2$ for the parameters $\mathbf{x} = (\mathbf{x}_1^T, \mathbf{x}_2^T)^T$, where $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$, $\bar{\mathbf{A}}_1 = \mathbf{P}_{\mathbf{A}_2}^\perp \mathbf{A}_1$

biases, the probability $P_{\text{HMD}}(\mathbf{b})$ will have a maximum for a certain bias (Fig. 24.11). With this approach one can thus evaluate whether the *worst-case* scenario

$$\max_{\mathbf{b}} P_{\text{HMD}}(\mathbf{b})$$

still satisfies ones criterion.

As the computations of the above approaches can be done without the need for having the actual measurements available, they are very useful for design verification purposes. Starting from a certain assumed design or measurement setup as described by \mathbf{A} and \mathbf{Q}_{yy} , one can then infer how well the design can be expected to be protected by the statistical testing against biases $\mathbf{b}_y = \mathbf{C}\mathbf{b}$.

Note that in the above considerations, we have taken the complete parameter vector \mathbf{x} into account. This need not be necessary, however, if only certain functions of \mathbf{x} , say $\boldsymbol{\theta} = \mathbf{F}^T \mathbf{x}$, are of interest for the application (e.g., only coordinates in the case of positioning, or only clocks in the case of time transfer). In that case one can follow the same approach as above, but now restricted to $\boldsymbol{\theta}$ [24.29, p. 110]. As a special case, consider the partitioning

$$\mathbf{x} = [\mathbf{x}_1^T, \mathbf{x}_2^T]^T$$

and assume that one is only interested in \mathbf{x}_1 (i. e., $\boldsymbol{\theta} = \mathbf{x}_1$ and $\mathbf{F}^T = [\mathbf{I}_{n_1}, \mathbf{0}]$). Then in analogy with (24.12),

$$\|\hat{\mathbf{x}}_1 - \mathbf{x}_1\|_{\mathbf{Q}_{\hat{\mathbf{x}}_1}}^2 \stackrel{\mathcal{H}_a}{\sim} \chi^2(n_1, \lambda_{\hat{\mathbf{x}}_1}^2),$$

with its noncentrality parameter

$$\lambda_{\hat{\mathbf{x}}_1}^2 = \mathbf{b}_{\hat{\mathbf{x}}_1}^T \mathbf{Q}_{\hat{\mathbf{x}}_1}^{-1} \mathbf{b}_{\hat{\mathbf{x}}_1}$$

linked to $\lambda_{\hat{y}}^2$ through the Pythagorean decomposition

$$\lambda_{\hat{y}}^2 = \lambda_{\hat{x}_1}^2 + \lambda_{\hat{x}_2|x_1}^2, \quad (24.55)$$

where

$$\lambda_{\hat{x}_2|x_1}^2 = \|\mathbf{P}_{\mathbf{A}_2} \mathbf{b}_y\|_{\mathbf{Q}_{yy}}^2.$$

In geometric terms, $\lambda_{\hat{x}_1}$ is related to $\lambda_{\hat{y}}$ and λ_y as (Fig. 24.12)

$$\begin{aligned} \lambda_{\hat{x}_1} &= \lambda_{\hat{y}} \cos(\phi_1) \\ &= \lambda_y \cos(\phi) \cos(\phi_1). \end{aligned} \quad (24.56)$$

Thus now also the angle ϕ_1 plays a role in determining how much of the actual bias \mathbf{b}_y is passed on to the LS solution of \mathbf{x}_1 . The larger the angle ϕ_1 , the smaller the impact.

Example 24.12 Example 3 continued

Under the assumptions of Example 24.3, the influential BNR (24.13) for \mathbf{x}_1 of $\mathbf{x} = [\mathbf{x}_1^T, \mathbf{x}_2^T]^T$ is

$$\lambda_{\hat{\mathbf{x}}_1} = \left(\frac{\sigma_{\hat{y}_i}^2 - \sigma_{\hat{y}_i|x_1}^2}{\sigma_{y_i}^2} \right)^{\frac{1}{2}} \frac{|b|}{\sigma_{y_i}}. \quad (24.57)$$

Hence, a large reduction in the BNR takes place if the constraining of \mathbf{x}_1 would not do much for the precision improvement of the adjusted observations. ■

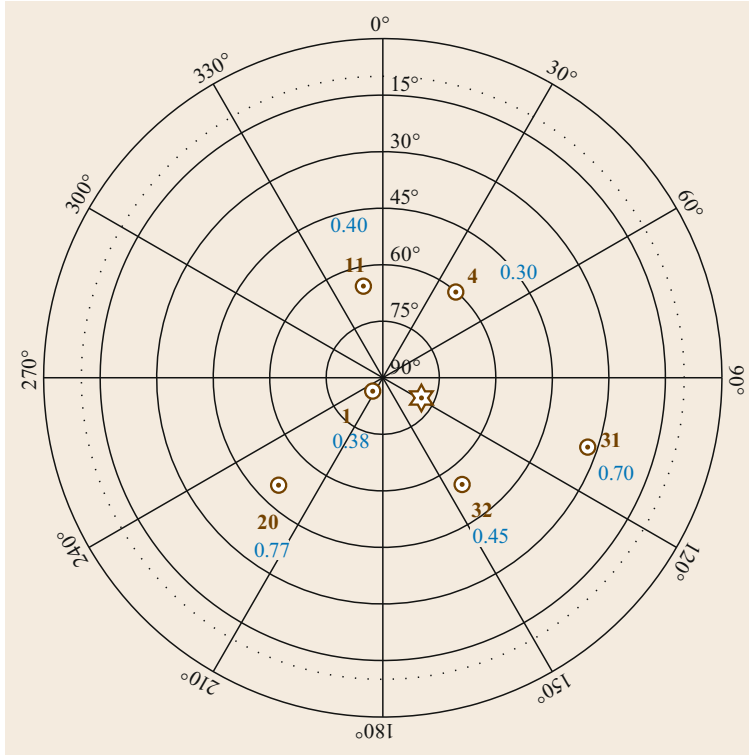


Fig. 24.13 Skyplot of PRNs 1, 4, 11, 20, 31 and 32, showing their values $c_i^T \mathbf{P}_G c_i$ in blue. The direction of the average $\mathbf{g}_c = \frac{1}{m} \mathbf{G}^T \mathbf{e}$ is shown as a star

Example 24.13 Example 24.7 continued

Application of (24.55) to the model of Example 24.7 allows one to study the impact on the intercept and slope separately. The influential BNR of the *intercept*, $\lambda_{\hat{x}_1} = |b_{\hat{x}_1}|/\sigma_{\hat{x}_1}$, reads

$$\lambda_{\hat{x}_1} = \left(\frac{1 - \frac{a_i^2}{\sum_{j=1}^m a_j^2}}{1 - \frac{\bar{a}_i^2}{\sum_{j=1}^m \bar{a}_j^2} - \frac{1}{m}} - 1 \right)^{\frac{1}{2}} \lambda_{\hat{e}}. \quad (24.58)$$

Thus $\lambda_{\hat{x}_1}$ is small if a_i is large and/or \bar{a}_i is small.

Similarly the influential BNR of the *slope*, $\lambda_{\hat{x}_2} = |b_{\hat{x}_2}|/\sigma_{\hat{x}_2}$, reads

$$\lambda_{\hat{x}_2} = \left(\frac{1}{1 - \frac{\frac{m}{m-1} \bar{a}_i^2}{\sum_{j=1}^m \bar{a}_j^2}} - 1 \right)^{\frac{1}{2}} \lambda_{\hat{e}}. \quad (24.59)$$

Thus $\lambda_{\hat{x}_2} = 0$ if $\bar{a}_i = 0$. Hence the effect of a possible nondetected outlier the size of the MDB on the *slope* estimator \hat{x}_2 is insignificant if a_i is close to a_c . ■

Example 24.14 Pruning of satellites

Consider a code-based single receiver model having (24.48) as design matrix and

$$\mathbf{Q}_{yy} = \sigma^2 \mathbf{I}_m$$

as variance matrix. With the partitioned parameter vector $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2^T)^T$, in which the first entry is the clock and the second contains the three E-N-U coordinate increments, $\mathbf{x}_2 = (E, N, U)^T$, the influential position BNR for an outlier in the i th code observable reads

$$\lambda_{\hat{x}_{2(i)}} = \left(\frac{c_i^T \mathbf{P}_G c_i}{1 - \left(\frac{1}{m} + c_i^T \mathbf{P}_G c_i \right)} \right)^{\frac{1}{2}} \lambda_{\hat{e}}. \quad (24.60)$$

This BNR can now be used to measure the positional significance of MDB-sized code outliers. The larger the ratio $\lambda_{\hat{x}_{2(i)}}/\lambda_{\hat{e}}$, i.e., the larger the scalar $c_i^T \mathbf{P}_G c_i$, the more significant a code outlier of MDB-size is for positioning. The values $c_i^T \mathbf{P}_G c_i$ are given in Fig. 24.13 for each of the satellites shown. The direction of the average \mathbf{g}_c is also shown. ■

24.4 Testing Procedure

24.4.1 Detection, Identification and Adaptation

So far we considered the testing of the null hypothesis \mathcal{H}_0 against one particular alternative hypothesis \mathcal{H}_a . In most practical applications however, it is usually not only one single mismodeling error one is concerned about, but quite often many more than one. This implies that one needs a *testing procedure* for handling the various alternative hypotheses \mathcal{H}_{a_i} . In this subsection we discuss a way of structuring such a testing procedure. The detection, identification and adaptation (DIA) procedure consists of the following three steps:

1. *Detection*: An overall model test is performed to diagnose whether an unspecified model error occurred.
2. *Identification*: After detection of a model error, identification of the potential source of model error is needed.
3. *Adaptation*: After identification of a model error, adaptation of the null hypothesis is needed to reduce the presence of biases in the solution.

Detection

This first step consists of a check on the overall validity of the null hypothesis \mathcal{H}_0 . It provides information on whether one can have any confidence in the assumed null hypothesis, without the explicit need to specify any particular alternative hypothesis. This implies that one opposes the null hypothesis to the most relaxed alternative hypothesis possible, namely $\mathcal{H}_a : E(\mathbf{y}) \in \mathbb{R}^m$. Under this hypothesis no restrictions are imposed on the observables. The test makes use of the statistic T_{m-n} (24.33) and reads

$$\text{Reject } \mathcal{H}_0 \text{ if } T_{m-n} = \|\hat{\mathbf{e}}\|_{\mathbf{Q}_{yy}}^2 > \chi_\alpha^2(m-n, 0) \quad (24.61)$$

and accept (not reject) otherwise.

The detection step constitutes a safeguard against all possible types of modeling errors. If the test is passed, the system can be considered available and no further action is required. If the test is rejected one may either consider the system unavailable or move on to the identification step. The latter is done if one has a fair idea of the type of modeling errors that may occur.

Identification

Assuming that the detection step led to a rightful rejection of the null hypothesis, one can try to search for the likely model misspecifications. That is, one can then try to identify the model error that caused the rejection of the null hypothesis. This implies that one will have to

specify, through the matrix \mathbf{C} , the type of likely model error $\mathbf{b}_y = \mathbf{C}\mathbf{b}$. This specification of possible alternative hypotheses is application-dependent and is one of the more difficult tasks in hypothesis testing. It depends on experience, which type of model errors are considered likely.

One can have alternative hypotheses of varying degrees of freedom q , as well as more than one with the same degrees of freedom. Let us first assume that the degrees of freedom of all alternatives are the same and equal to q . The hypotheses considered are then

$$\mathcal{H}_{q,i} : \mathbf{b}_y = \mathbf{C}_i \mathbf{b}_i, \mathbf{b}_i \in \mathbb{R}^q, i = 1, \dots, r_q. \quad (24.62)$$

The identified hypothesis \mathcal{H}_{q,j_q} is then the one for which

$$j_q = \arg \max_i T_{q,i}. \quad (24.63)$$

Thus it is the maximum value of the r_q test statistics $T_{q,i}$ that is singled out as the most likely alternative hypothesis of q degrees of freedom. This can also be understood by considering the relation (24.31)

$$\|\hat{\mathbf{e}}_{q,i}\|_{\mathbf{Q}_{yy}}^2 = \|\hat{\mathbf{e}}\|_{\mathbf{Q}_{yy}}^2 - T_{q,i}.$$

It shows that selecting the largest $T_{q,i}$ amounts to selecting the *best-fitting* model among the $\mathcal{H}_{q,i}$ s with q degrees of freedom, i. e., the one with smallest $\|\hat{\mathbf{e}}_{q,i}\|_{\mathbf{Q}_{yy}}^2$ for $i = 1, \dots, r_q$.

This simple approach of choosing the hypothesis with largest value of its test statistic does not work anymore in the case where (24.63) is applied to hypotheses with varying degrees of freedom. Since $E(T_{q,i} | \mathcal{H}_0) = q$, the mean of the test statistic gets larger for larger q . Hence, if identification would be based on the size of $T_{q,i}$, one would tend to select under \mathcal{H}_0 hypotheses with largest degrees of freedom. Although one could compensate by considering the maximum of $T_{q,i} - q$ or $T_{q,i}/q$ instead, it is better to base the identification on test statistics that have the same distribution under the null hypothesis.

A straightforward way to transform different random variables such that their transformations have the same distribution is by using their cumulative distribution function (CDF) as the transformation. If a and b are two random variables with CDFs $F_a(x)$ and $F_b(x)$, then $F_a(a)$ and $F_b(b)$ have the same distribution, and so have $1 - F_a(a)$ and $1 - F_b(b)$. Their common distribution is then the uniform distribution on the interval $[0, 1]$ [24.60]. Thus if we are using normally distributed

w -test statistics and chi-squared distributed test statistics T_q with varying degrees of freedom q , we can transform the latter so that they also become standard normally distributed under \mathcal{H}_0 . This transformation is given as $w(x) = \Phi^{-1}(\chi_q(x))$, with $\chi_q(x)$ and $\Phi(x)$ the CDFs of $\chi^2(q, 0)$ and $\mathcal{N}(0, 1)$ respectively. As a result we have

$$w(T_q) \stackrel{\mathcal{H}_0}{\sim} \mathcal{N}(0, 1) .$$

In the above F -transformations we recognize the transformation to the p -values of a and b respectively. Hence, instead of transforming back to the normal distribution, one can in the general case of varying degrees of freedom also use the concept of the uniformly distributed p -values and select the hypothesis with smallest $p(T_{q,i})$ (24.25) as the most likely hypothesis,

$$\min_{q,i} p(T_{q,i}) . \quad (24.64)$$

These p -values do not need to be compared to a criterion if it is assumed that the set (24.62) indeed covers all hypotheses that can possibly occur. The identification can then be restricted to choosing the most likely alternative. However, if one leaves the option open of alternative hypotheses that are not in the set, then a comparison with a chosen level of significance is needed so as to decide whether the most likely alternative is indeed likely enough. An alert is then given if none are considered likely enough. In that case we have rejection at the detection stage, but no identification among the set of specified alternative hypotheses. Different approaches exist to link the levels of significance of identification to that of detection; see for example, [24.34, 61–63].

Adaptation

Once one or more likely model errors have been identified, a corrective action needs to be undertaken in order to get the null hypothesis accepted. Here, one of the two following approaches can be used in principle. Either one replaces the data or part of the data with new data such that the null hypothesis does get accepted, or, one replaces the original null hypothesis with a new hypothesis that does take the identified model errors into account. The first approach amounts to a remeasurement of (part of) the data. In the second approach no remeasurement is undertaken. Instead the identified hypothesis becomes the new null hypothesis. Thus the model of the null hypothesis is enlarged by adding the additional parameters as identified in the identification step. With \mathcal{H}_a being the identified hypothesis and $\hat{\mathbf{b}}$ its estimated bias vector, the adaptation amounts to cor-

recting the biased solution $\hat{\mathbf{x}}$ through

$$\hat{\mathbf{x}}_a = \hat{\mathbf{x}} - \mathbf{A}^+ \mathbf{C} \hat{\mathbf{b}} \stackrel{\mathcal{H}_a}{\sim} \mathcal{N}(\mathbf{x}, \mathbf{Q}_{\hat{\mathbf{x}}_a \hat{\mathbf{x}}_a}) , \quad (24.65)$$

with $\mathbf{Q}_{\hat{\mathbf{x}}_a \hat{\mathbf{x}}_a} = \mathbf{Q}_{\hat{\mathbf{x}} \hat{\mathbf{x}}} + \mathbf{A}^+ \mathbf{C} \mathbf{Q}_{\hat{\mathbf{b}} \hat{\mathbf{b}}} \mathbf{C}^T \mathbf{A}^{+\top}$. Note that although $\hat{\mathbf{x}}_a$ is unbiased, it has a poorer precision than the biased solution $\hat{\mathbf{x}}$. This is the price one pays for reducing the bias from the solution.

Once the adaptation step is completed, one of course still has to make sure whether the newly created situation is acceptable or not. This at least implies a repetition of the detection step. When adaptation is applied, one also has to be aware of the fact that since the model may have changed, also the *strength of the model* may have changed. In fact, when the model is adapted through the addition of more explanatory parameters, the model has become weaker in the sense that the test statistics will now have less detection and identification power. It depends on the particular application at hand whether this is considered acceptable or not.

24.4.2 Data Snooping

An important example of multiple hypothesis testing is the problem of screening the observations for possible outliers [24.34, 56, 64–73].

The set of alternative hypotheses will then include those that describe outliers in individual observations. Here we restrict ourselves to the case of one outlier at a time. In that case there are as many alternative hypotheses as there are observations, with each alternative describing the outlier in an observation. Thus

$$\mathcal{H}_i : \mathbf{b}_y = \mathbf{c}_i b_i, \quad b_i \in \mathbb{R} \quad , \quad i = 1, \dots, m , \quad (24.66)$$

with \mathbf{c}_i the canonical unit vector having 1 as its i th entry. Thus b_i is the scalar outlier and i indicates the observation that is assumed to be affected by it. By letting i run from 1 up to and including m , one can screen the whole dataset for the presence of potential blunders in the individual observations. The test statistic w_i ((24.35), with $\mathbf{c} := \mathbf{c}_i$), which returns the in absolute value largest value, then pinpoints the observation that is most likely corrupted with an outlier or blunder. Its significance is measured by comparing the value of the test statistic with the critical value k . Thus, in analogy with (24.63), the j th observation is suspected to have an outlier, when

$$\begin{aligned} |w_j| &= \max_i |w_i| \geq k \quad , \quad \text{or} \\ p_j &= \min_i p_i \leq \alpha \quad , \end{aligned} \quad (24.67)$$

with critical value

$$k = \mathcal{N}_{\frac{1}{2}\alpha}^-(0, 1) ,$$

p -value $p_i = 2[1 - \Phi(|w_i|)]$ and $\Phi(x)$ the standard normal distribution function. This procedure of screening each individual observation for the presence of an outlier is known as *data snooping* [24.34, 61]. In the special case of a diagonal variance matrix \mathbf{Q}_{yy} it is sometimes also referred to as the *maximum residual technique* [24.27]. The adaptation step of data snooping can follow (24.65). Due to the special structure of $\mathbf{C} = \mathbf{c}_i$, this adaptation step in essence means that $\hat{\mathbf{x}}_a$ is computed with the measurement corresponding to the maximum w -statistic excluded.

Although formulation (24.66) assumes the presence of a single outlier, (iterated) data snooping is also used to find multiple outliers. For a discussion on multiple outlier testing, see for example [24.74–82]. Also note that one can generalize the constant critical value k of (24.67) to a variable setting, i.e., a critical value k_i per w_i -test, thus allowing for extra flexibility in testing.

When evaluating a multiple testing procedure, one can in principle not work anymore with the binary probabilities of false alarm and missed detection of (24.24). That is, for a consideration of the error probabilities, one has to consider the complete partitioning of the multidimensional sample space into acceptance and rejection regions. In the case of the above data snooping with constant critical value k this implies that the acceptance region is given by the origin centered hyper-cube

$$\mathcal{A} = \{x \in \mathbb{R}^m \mid |x_i| \leq k, i = 1, \dots, m\}.$$

Defining the m -vector of w_i -statistics as

$$\mathbf{w} = (w_1, \dots, w_m)^T,$$

the probabilities of false alarm and missed detection now read

$$\begin{aligned} P_{\text{FA}} &= P(\mathbf{w} \notin \mathcal{A} \mid \mathcal{H}_0), \\ P_{\text{MD}_i} &= P(\mathbf{w} \in \mathcal{A} \mid \mathcal{H}_i). \end{aligned} \quad (24.68)$$

These probabilities can be computed using Monte Carlo integration [24.60]. The binary missed detection probability $P(|w_i| \leq k \mid \mathcal{H}_i)$ is an upper bound for P_{MD_i} . Thus if the binary probability of missed detection is acceptable so is the overall one. This is however not true for the false alarm. To get an idea of the multidimensionality effect, assume that the m number of w_i -test statistics is independent and that all m tests are executed with a level of significance α . Then under \mathcal{H}_0 , $(1 - \alpha)^m$ is the probability that all tests are accepted and thus $P_{\text{FA}} = 1 - (1 - \alpha)^m$. Hence, with a choice $\alpha = 5\%$, this would already give a false alarm probability of $P_{\text{FA}} = 64\%$ if $m = 20$. This shows that care has to be taken in choosing the level of significance for the individual tests. The probability becomes an upper bound

of the false alarm if the assumption of independence is dropped

$$P_{\text{FA}} \leq 1 - (1 - \alpha)^m \approx m\alpha. \quad (24.69)$$

The approximation of the upper bound is sharp for small α . This result can now be used to control the false alarm rate of data snooping. If one requires the overall false alarm rate of data snooping (DS) to be not greater than $P_{\text{FA}} = \alpha_{\text{DS}}$, then the level of significance for each individual test could be chosen as $\alpha = 1 - (1 - \alpha_{\text{DS}})^{\frac{1}{m}} \approx \alpha_{\text{DS}}/m$. The latter correction is known as the Bonferroni correction [24.83].

Although the above level-of-significance adjustment certainly controls the overall false alarm rate of data snooping, it may sometimes, when m is large, also lead to too small values and therefore to larger probabilities of missed detection. This is due to the fact that the above upper bound discards the dependence among the w_i -test statistics. There do exist less conservative procedures, including those that take a somewhat different criterion for the multidimensional false alarm. The approach above is based on controlling the probability that one or more of the rejected hypotheses is true. An alternative is the false discovery rate approach. Instead of controlling the probability that one or more of the rejected hypotheses is true, it controls the *expected* number of false rejections from among the rejected hypotheses; see for example [24.63, 84–86].

The consequence of the multiple testing involved in data snooping is that for the evaluation of the probability of hazardous missed detection, one has next to the missed detection probability, P_{MD_i} (24.68), also a corresponding probability of hazardous occurrence, $P_{\text{H}_i} = P(P_n \geq \chi_n^2(n, 0) \mid \mathcal{H}_i)$, for each alternative hypothesis \mathcal{H}_i . Hence, we have per alternative hypothesis, $P_{\text{HMD}}(b_i) = P_{\text{H}_i} \times P_{\text{MD}_i}$. These conditional probabilities (i.e., conditioned on \mathcal{H}_i) can be evaluated individually, or unconditionally as a weighted average,

$$P_{\text{HMD}}(\mathbf{b}) = \sum_{i=1}^m p_i P_{\text{HMD}}(b_i), \quad (24.70)$$

with the weights summing up to one, $\sum_{j=1}^m p_j = 1$. If the a priori probabilities for the occurrence of the individual alternative hypotheses, $P(\mathcal{H}_i)$ $i = 1, \dots, m$ are available, then the weights are given by their relative frequencies,

$$p_i = P(\mathcal{H}_i) / \sum_{j=1}^m P(\mathcal{H}_j).$$

In case one would like to include the hazardous occurrence under \mathcal{H}_0 as well in the evaluation, then (24.70)

Table 24.2 Outliers b (m) in PRN pseudoranges, epochs of their occurrence k and their estimated values \hat{b} (m)

Epoch	PRN	b	\hat{b}
$k = 50$	22	10	9.5
$k = 100$	31	-10	-9.3
$k = 150$	19	10	9.4
$k = 200$	17	5	5.1
$k = 250$	23	2	2.1
$k = 300$	31	2	2.2

needs extension with the probability $P_0 = \eta \times (1 - \alpha)$, to give the probability of hazardous misleading information $P_{\text{HMI}}(\mathbf{b}) = p_0 P_0 + (1 - p_0) P_{\text{HMD}}(\mathbf{b})$, with p_0 being the a priori occurrence probability of the null hypothesis [24.87]. Since the relative occurrences of the alternative hypotheses \mathcal{H}_i (24.66) are often easier to specify than the probability of occurrence of the null hypothesis itself, it is usually easier to work with (24.70). As both probabilities show the same variability as function of the biases, both have the same *worst-case bias solution*.

Example 24.15 Differential GPS

We now briefly discuss a Differential Global Positioning System (DGPS) testing example taken from [24.57]. The model on which this example is based is one of instantaneous relative positioning using single frequency code data. Both receivers tracked the same seven satellites every epoch using a sampling

interval of ten seconds. Atmospheric delays were neglected because of the short baseline length (≈ 3 km). The standard deviation of the undifferenced code observable was set at $\sigma_p = 30$ cm in zenith. The skyplot over the observation period is shown in Fig. 24.14. At six different epochs over the observation period, one of the seven pseudoranges was artificially corrupted by an outlier. The epoch number (k), the corrupted pseudorange (PRN) and the size of the outlier (b) are shown in the first three columns of Table 24.2. The last column shows the least-squares estimate \hat{b} of the outlier.

Over the observation period the MDBs were about 2–3 m. The results of the detection step are depicted in Fig. 24.15. Along the vertical axis the normalized test statistic

$$\frac{T_{q=3}}{\chi^2_{\alpha}(3, 0)}$$

is shown. The null hypothesis is rejected when this value exceeds one. The figure clearly shows that model errors are indeed detected at the above given six different epochs. Identification of the model errors is therefore needed for these six epochs. For each epoch this is done with the identification test statistic w_i . For each epoch, there are seven of such test statistics, one for each (single differenced) pseudorange. The values of these test statistics are shown in Table 24.3. The in

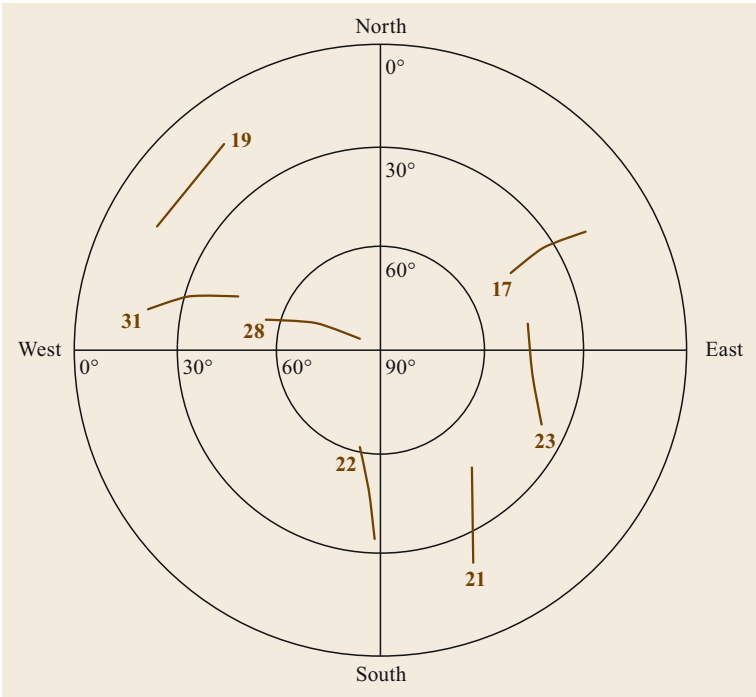


Fig. 24.14 Skyplot of seven GPS satellites with their PRN numbers; (after [24.4])

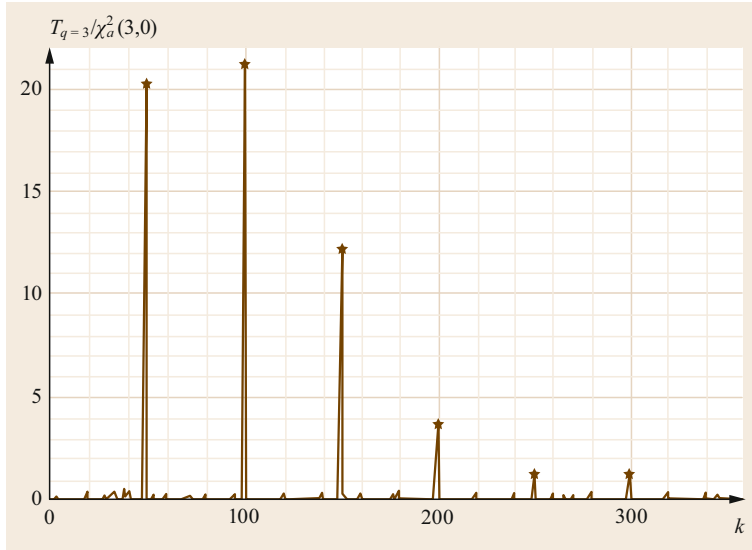


Fig. 24.15 Time series of normalized test statistic $T_{q=3}/\chi^2_{\alpha}(3,0)$ (after [24.4])

Table 24.3 Data snooping w_i -values ($i = 1, \dots, 7$) for the seven pseudoranges at the six epochs $k = 50, 100, 150, 200, 250, 300$

PRN/ k	50	100	150	200	250	300
17	1.9	-4.6	-6.4	6.8	-3.4	0.2
19	7.5	-14.8	12.4	-3.9	0.6	-3.1
21	-7.5	5.5	-1.8	-0.1	-0.5	0.3
22	15.9	3.5	3.7	2.5	-1.0	-0.7
23	-2.7	-4.0	2.5	-6.0	3.9	0.7
31	-2.2	-16.3	-11.4	1.7	1.1	3.9
28	-13.9	6.8	3.5	-0.4	-2.7	-3.6

absolute value largest test statistics are shown in *italics*. They indicate the pseudoranges that are most likely corrupted by outliers. The values in *italic* all exceed the critical value $\mathcal{N}_{\alpha/2}(0, 1) = 3.29$. In this case all code outliers were correctly identified; compare Tables 24.2 and 24.3.

24.4.3 Unknowns in the Stochastic Model

So far we assumed the variance matrix of the observables to be known. In some applications it may happen however that the variance matrix \mathbf{Q}_{yy} is only partly known. The simplest of such cases occurs if the variance matrix is known up to an unknown scale factor σ^2 , the so-called *variance of unit weight*. Then

$$\mathbf{Q}_{yy} = \sigma^2 \mathbf{C}_{yy},$$

with known cofactor matrix \mathbf{C}_{yy} , but unknown scalar σ^2 . If this is the case, then none of the previously discussed test statistics can be computed, since (24.30)

$$T_q = \frac{\|\hat{\mathbf{b}}\|_{\mathbf{C}_{bb}}^2}{\sigma^2}. \quad (24.71)$$

To cope with this situation, the idea is to replace the unknown σ^2 with an unbiased estimator of it. There are two such candidates, namely

$$\hat{\sigma}^2 = \frac{\|\hat{\mathbf{e}}\|_{\mathbf{C}_{yy}}^2}{m-n} \quad \text{and} \quad \hat{\sigma}_a^2 = \frac{\|\hat{\mathbf{e}}_a\|_{\mathbf{C}_{yy}}^2}{m-n-q}. \quad (24.72)$$

Thus if σ^2 is unknown, one can work, instead of with (24.71), with either one of the following two test statistics

$$T'_q = \frac{\|\hat{\mathbf{b}}\|_{\mathbf{C}_{bb}}^2}{q\hat{\sigma}^2}, \quad T''_q = \frac{\|\hat{\mathbf{b}}\|_{\mathbf{C}_{bb}}^2}{q\hat{\sigma}_a^2}. \quad (24.73)$$

The statistic T''_q has an F -distribution with q and $(m-n-q)$ degrees of freedom, while $q/(m-n)T'_q$ has a beta distribution with $q/2$ and $(m-n-q)/2$ degrees of freedom [24.31, 88]. The two test statistics are functionally related as

$$T'_q = \frac{(m-n)T''_q}{(m-n-q) + qT''_q}. \quad (24.74)$$

Note that they cannot be used for

$$q = m - n ,$$

since then $T'_q = 1$ and T''_q is undefined. Hence, no detection test (24.61) exists when σ^2 is unknown. For the one-dimensional case $q = 1$, we have

$$\begin{aligned} T'_{q=1} &= (w')^2 \quad \text{and} \\ T''_{q=1} &= (w'')^2 , \end{aligned} \quad (24.75)$$

with

$$w' = \frac{\sigma}{\hat{\sigma}} w , \quad w'' = \frac{\sigma}{\hat{\sigma}_a} w . \quad (24.76)$$

Thus in the case where σ^2 is unknown, w' or w'' can be used instead of w . They have a τ - and Student t -distribution respectively [24.64, 89]. In the case where \mathbf{C}_{yy} is diagonal and \mathbf{c} is the canonical unit-vector, w' and w'' are referred to as the *internal* and *external* Studentized residuals respectively [24.65, 70].

24.5 Recursive Model Validation

24.5.1 Model and Filter

In the previous section the parameter vector \mathbf{x} was assumed to be time invariant and completely unknown. In the present section we will relax these assumptions. First, the parameter vector \mathbf{x} (from now on referred to as the state vector) will be allowed to vary with time. Thus instead of the notation \mathbf{x} , we will now use the notation \mathbf{x}_k to indicate the (discrete) time epoch k to which \mathbf{x} refers. Secondly, we will assume that the parameter vector, instead of being deterministic, is now a random vector as well.

The model on which the filter is based consists of two parts (Chap. 22). As before we have an observational model that links the observables to the state vector. But in addition we now also have a dynamic model, describing the variability of the state vector over time. Both the measurement model and dynamic model constitute the null hypothesis \mathcal{H}_0 .

Measurement model: The link between the vector of observables $\mathbf{y}_k \in \mathbb{R}^{m_k}$ and the state-vector $\mathbf{x}_k \in \mathbb{R}^n$ is assumed given as

$$\mathbf{y}_k = \mathbf{A}_k \mathbf{x}_k + \mathbf{n}_k, \quad k = 0, \dots, \quad (24.77)$$

with

$$\begin{aligned} E(\mathbf{x}_0) &= \bar{\mathbf{x}}_0 , \\ E(\mathbf{n}_k) &= \mathbf{0} , \\ C(\mathbf{x}_0, \mathbf{n}_k) &= \mathbf{0} , \quad \text{and} \\ C(\mathbf{n}_k, \mathbf{n}_l) &= \mathbf{R}_k \delta_{k,l} , \end{aligned}$$

where $C(\mathbf{u}, \mathbf{v})$ denotes the covariance between two random vectors \mathbf{u} and \mathbf{v} , \mathbf{R}_k is the variance matrix of the measurement noise \mathbf{n}_k and $\delta_{k,l}$ is the Kronecker delta. Thus the zero-mean measurement noise \mathbf{n}_k is assumed to be uncorrelated in time and to be uncorrelated with the initial state-vector \mathbf{x}_0 .

Dynamic model: The linear dynamic model, describing the time evolution of \mathbf{x}_k , is given as

$$\mathbf{x}_k = \Phi_{k,k-1} \mathbf{x}_{k-1} + \mathbf{d}_k, \quad k = 1, \dots, \quad (24.78)$$

with

$$\begin{aligned} E(\mathbf{d}_k) &= \mathbf{0} , \\ C(\mathbf{x}_0, \mathbf{d}_k) &= \mathbf{0} , \\ C(\mathbf{d}_k, \mathbf{n}_l) &= \mathbf{0} , \\ C(\mathbf{d}_k, \mathbf{d}_l) &= \mathbf{S}_k \delta_{k,l} , \end{aligned}$$

where $\Phi_{k,k-1}$ denotes the transition matrix and \mathbf{S}_k the variance matrix of the system noise $\mathbf{d}_k \in \mathbb{R}^n$. The system noise \mathbf{d}_k is thus also assumed to have a zero mean, to be uncorrelated in time and to be uncorrelated with the initial state-vector and the measurement noise.

If the mean of the initial state, $E(\mathbf{x}_0)$, is assumed unknown, it first needs to be estimated. Since we will assume here that the initial state $\hat{\mathbf{x}}_{0|0}$ and its error-variance matrix $\mathbf{P}_{0|0}$ are known, we start the recursion from these initial values. The recursive procedure consists then of two steps for every epoch, a time update (TU) and a measurement update (MU)

$$\begin{aligned} \hat{\mathbf{x}}_{k|k-1} &= \Phi_{k,k-1} \hat{\mathbf{x}}_{k-1|k-1} , \\ \hat{\mathbf{x}}_{k|k} &= \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k (\mathbf{y}_k - \mathbf{A}_k \hat{\mathbf{x}}_{k|k-1}) . \end{aligned} \quad (24.79)$$

The time update propagates the state vector estimate of epoch $k-1$ to the next epoch k , thus giving the predicted state vector of epoch k . The measurement update combines in a least-squares sense the predicted state vector with the observations of epoch k to produce the filtered state vector at that epoch. The above pair of equations is often referred to as the Kalman filter and matrix \mathbf{K}_k as the Kalman gain matrix [24.90–93]. The (error) variance matrices of the predicted and filtered state vector

are given as

$$\begin{aligned}\mathbf{P}_{k|k-1} &= \Phi_{k,k-1} \mathbf{P}_{k-1|k-1} \Phi_{k,k-1}^\top + \mathbf{S}_k, \\ \mathbf{P}_{k|k} &= (\mathbf{I}_n - \mathbf{K}_k \mathbf{A}_k) \mathbf{P}_{k|k-1}.\end{aligned}\quad (24.80)$$

Due to the presence of the system's noise, we usually have a worsening of the precision in the time-update step, $\mathbf{P}_{k|k-1} > \mathbf{P}_{k-1|k-1}$, while in the measurement-update, with the additional measurements, the precision improves again such that $\mathbf{P}_{k|k} < \mathbf{P}_{k|k-1}$.

An important role in the process of recursive model validation is played by the so-called predicted residuals. They play a role that is similar to that of the least-squares residuals in the previous sections. The predicted residual and its variance matrix are given as

$$\begin{aligned}\mathbf{v}_k &= \mathbf{y}_k - \mathbf{A}_k \hat{\mathbf{x}}_{k|k-1}, \\ \mathbf{Q}_{\mathbf{v}_k \mathbf{v}_k} &= \mathbf{R}_k + \mathbf{A}_k \mathbf{P}_{k|k-1} \mathbf{A}_k^\top.\end{aligned}\quad (24.81)$$

The predicted residuals have the important property that they are zero-mean and uncorrelated in time

$$E(\mathbf{v}_k) = \mathbf{0}, \quad C(\mathbf{v}_k, \mathbf{v}_l) = \mathbf{Q}_{\mathbf{v}_k \mathbf{v}_l} \delta_{k,l}. \quad (24.82)$$

It are these properties that make it possible to formulate a recursive testing procedure that can be executed in parallel to the Kalman filter [24.10, 25, 57, 75, 94, 95],

24.5.2 Models and UMPI Test Statistic

The above Kalman filter produces optimal estimators of the state vector with well-defined statistical properties. This optimality is only guaranteed however as long as the assumptions underlying the model, i. e., the null hypothesis \mathcal{H}_0 , hold. Misspecifications in \mathcal{H}_0 will invalidate the results of estimation and thus also any conclusions based on them. In the following we restrict attention, as before, to misspecifications in the mean. This implies that the predicted residuals have a zero-mean under \mathcal{H}_0 , but not so under \mathcal{H}_a . Thus if we collect all k vectors of predicted residuals in one vector $\mathbf{v} = (\mathbf{v}_1^\top, \dots, \mathbf{v}_k^\top)^\top$, the null and alternative hypotheses can be expressed in terms of the predicted residuals as

$$\mathcal{H}_0 : E(\mathbf{v}) = \mathbf{0} \text{ versus } \mathcal{H}_a : E(\mathbf{v}) = \mathbf{C}_v \mathbf{b}. \quad (24.83)$$

The \mathbf{C}_v -matrix is of order $\sum_{i=1}^k m_i \times q$ and follows from propagating the assumed misspecifications in the mean through the time update and measurement update respectively.

As before one can now estimate \mathbf{b} under \mathcal{H}_a and test its significance using the UMPI test statistic

$$T_q = \hat{\mathbf{b}}^\top \mathbf{Q}_{\hat{\mathbf{b}} \hat{\mathbf{b}}}^{-1} \hat{\mathbf{b}} \stackrel{\mathcal{H}_0}{\sim} \chi^2(q, 0) \quad (24.84)$$

in which the bias estimator $\hat{\mathbf{b}}$ and its variance matrix $\mathbf{Q}_{\hat{\mathbf{b}} \hat{\mathbf{b}}}$ are computed from the vector of predicted residuals \mathbf{v} as

$$\begin{aligned}\hat{\mathbf{b}} &= (\mathbf{C}_v^\top \mathbf{Q}_{\mathbf{v} \mathbf{v}}^{-1} \mathbf{C}_v)^{-1} \mathbf{C}_v^\top \mathbf{Q}_{\mathbf{v} \mathbf{v}}^{-1} \mathbf{v}, \\ \mathbf{Q}_{\hat{\mathbf{b}} \hat{\mathbf{b}}} &= (\mathbf{C}_v^\top \mathbf{Q}_{\mathbf{v} \mathbf{v}}^{-1} \mathbf{C}_v)^{-1}.\end{aligned}\quad (24.85)$$

Using this result one can express the test statistic (24.84) also directly in the predicted residuals

$$T_q = \mathbf{v}^\top \mathbf{Q}_{\mathbf{v} \mathbf{v}}^{-1} \mathbf{C}_v (\mathbf{C}_v^\top \mathbf{Q}_{\mathbf{v} \mathbf{v}}^{-1} \mathbf{C}_v)^{-1} \mathbf{C}_v^\top \mathbf{Q}_{\mathbf{v} \mathbf{v}}^{-1} \mathbf{v}. \quad (24.86)$$

Although this form is the equivalent of (24.32), it has the advantage of being expressed in the predicted residuals which, after all, are readily available at each measurement update (cf. (24.79) and (24.81)).

The above expressions form the basis for Kalman-filter model validation. They are however not yet in a form that facilitates *recursive* testing. This will become possible though if we make use of the block diagonal structure of the variance matrix of the vector of predicted residuals,

$$D(\mathbf{v}) = \mathbf{Q}_{\mathbf{v} \mathbf{v}} = \text{blockdiag}(\mathbf{Q}_{\mathbf{v}_1 \mathbf{v}_1}, \dots, \mathbf{Q}_{\mathbf{v}_k \mathbf{v}_k}). \quad (24.87)$$

With the use of this block diagonal structure, the test statistic (24.86) can be written as

$$\begin{aligned}T_q &= \left(\sum_{i=1}^k \mathbf{v}_i^\top \mathbf{Q}_{\mathbf{v}_i \mathbf{v}_i}^{-1} \mathbf{C}_{v_i} \right) \left(\sum_{i=1}^k \mathbf{C}_{v_i}^\top \mathbf{Q}_{\mathbf{v}_i \mathbf{v}_i}^{-1} \mathbf{C}_{v_i} \right)^{-1} \\ &\quad \times \left(\sum_{i=1}^k \mathbf{C}_{v_i}^\top \mathbf{Q}_{\mathbf{v}_i \mathbf{v}_i}^{-1} \mathbf{v}_i \right).\end{aligned}\quad (24.88)$$

It is this expression that we will use in the following as basis for the recursive detection, identification and adaptation. The design and performance analysis of recursive testing (e.g., MDB and BNR analyses) can be done in a way that is similar to what has been discussed earlier for the batch model; see for example [24.26, 35, 53, 96–98].

24.5.3 Local and Global Testing

We make a distinction between *local* model testing and *global* model testing (Fig. 24.16). We speak of local model testing when the tests performed at time k only depend on the predicted state vector at time k and the observations of time k . Thus for the local case we assume that no invalidation of the model has taken place

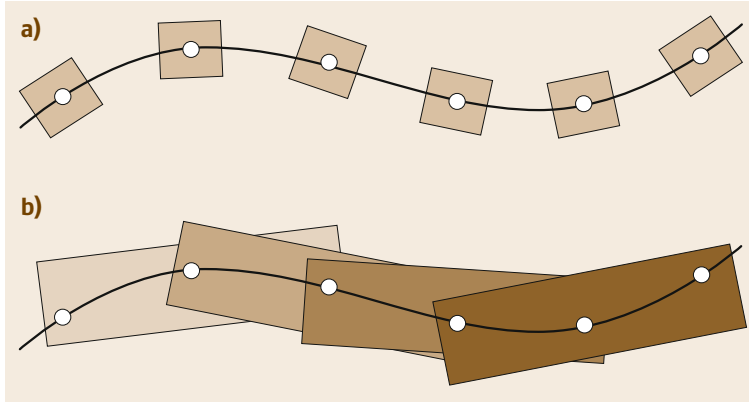


Fig. 24.16 Local (top) and global (bottom) testing (after [24.25])

prior to the present time of testing k . This implies that attention can be restricted to the following *local* hypotheses

$$\mathcal{H}_0^k : E(\mathbf{v}_k) = \mathbf{0} \text{ versus } \mathcal{H}_a^k : E(\mathbf{v}_k) = \mathbf{C}_{v_k} \mathbf{b} \quad (24.89)$$

in which the type of model error occurring at epoch k is specified through the $m_k \times q$ matrix \mathbf{C}_{v_k} .

If the test takes more than one epoch into account we speak of global testing. For those cases we consider the following *global* hypotheses

$$\mathcal{H}_0^{l,k} : E(\mathbf{v}^{l,k}) = \mathbf{0} \text{ versus } \mathcal{H}_a^{l,k} : E(\mathbf{v}^{l,k}) = \mathbf{C}_v^{l,k} \mathbf{b}, \quad (24.90)$$

with l the epoch the assumed model error is supposed to have started, k the current epoch,

$$\mathbf{v}^{l,k} = (\mathbf{v}_l^T, \dots, \mathbf{v}_k^T)^T,$$

and

$$\mathbf{C}_v^{l,k} = (\mathbf{C}_{v_l}^T, \dots, \mathbf{C}_{v_k}^T)^T.$$

This distinction between local and global is introduced in order to discriminate between model errors that have either a local or a more global character. By definition, the local procedure can be executed in real time. That is, the corrective action of the adaptation step is designed to coincide with the moment the model error occurred.

Global testing is of course more powerful than local testing. Hence, for certain model errors local testing may be too insensitive. Model errors that slowly build up as time proceeds may have a high probability of passing the local tests unnoticed. For such type of errors global testing is needed, due to their built-in memory

capabilities. The price paid for the higher power of the global tests is the delay between the time the model error started to occur and the time of detection. But such a delay will be acceptable if it is considered more important to detect the model error than to not detect it at all.

24.5.4 Recursive Detection

Local Detection

Local detection applies when one assumes the most relaxed version of the alternative in (24.89),

$$\mathcal{H}_0^k : E(\mathbf{v}_k) = \mathbf{0} \text{ versus } \mathcal{H}_a^k : E(\mathbf{v}_k) \in \mathbb{R}^{m_k}. \quad (24.91)$$

As no restrictions are imposed on the mean of the predicted residual under \mathcal{H}_a^k , the matrix \mathbf{C}_{v_k} in (24.89) is chosen as a square and regular matrix and thus $q = m_k$. Since $\mathbf{C}_{v_i} = \mathbf{0}$ for $i < k$, the corresponding test statistic for the *local overall model* (LOM) test follows from (24.88) as

$$T_{\text{LOM}}^k = \frac{\mathbf{v}_k^T \mathbf{Q}_{v_k v_k}^{-1} \mathbf{v}_k}{m_k} \stackrel{\mathcal{H}_0^k}{\sim} F(m_k, \infty, 0). \quad (24.92)$$

An unspecified model error is considered present at time k when $T_{\text{LOM}}^k > F_\alpha(m_k, \infty, 0)$. Note that we divided the quadratic form in (24.92) by the local redundancy m_k . This is not essential, but is merely done for practical reasons so as to have a graphical display of the time series of T_{LOM}^k fluctuate around the value of one, since

$$E(T_{\text{LOM}}^k | \mathcal{H}_0) = 1.$$

The normalization by m_k changes the distribution from a χ^2 -distribution into an F -distribution. A similar normalization is done in the case of the global detector.

Global Detection

The above LOM test may turn out to be too insensitive to detect global unmodeled trends. For those cases we consider the following *global* hypotheses

$$\mathcal{H}_0^{l,k} : E(\mathbf{v}^{l,k}) = \mathbf{0} \text{ versus } \mathcal{H}_a^{l,k} : E(\mathbf{v}^{l,k}) \in \mathbb{R}^{\sum_{i=l}^k m_i} . \quad (24.93)$$

Thus matrix $\mathbf{C}_v^{l,k}$ of (24.90) is chosen as a square and regular matrix and

$$q = \sum_{i=l}^k m_i .$$

The corresponding *global overall model* (GOM) test statistic can be written in recursive form as

$$T_{\text{GOM}}^{l,k} = T_{\text{GOM}}^{l,k-1} + g_k \left(T^k - m_k T_{\text{GOM}}^{l,k-1} \right) , \quad (24.94)$$

with

$$T^k = \mathbf{v}_k^T \mathbf{Q}_k^{-1} \mathbf{v}_k$$

and the scalar gain (Fig. 24.17)

$$g_k = \frac{1}{\sum_{i=l}^k m_i} .$$

The recursion is initialized with $T_{\text{GOM}}^{l,l} = T_{\text{LOM}}^l$. An unspecified model error within the interval $[l, k]$ is considered detected at time k when

$$T_{\text{LOM}}^{l,k} > F_\alpha \left(\sum_{i=l}^k m_k, \infty, 0 \right) .$$

Note that this test reduces to its local counterpart when $l = k$.

A practical problem with the above test is the choice of l , the time that the model error is assumed to have started to occur. Since the starting time of the model error is unknown a priori, one has to start in principle

with $l = 1$. A fixed value of l however implies a growing memory recursion, with the practical problem of a possible long delay in time of detection. Rejection of H_0 at time k with the GOM test may imply namely that a global model error started to occur as early as time $l = 1$. In order to reduce the time of delay, it is worthwhile to consider a moving window of length N by constraining l to $k - N + 1 \leq l \leq k$. When choosing N one of course has to make sure that the detection power of the test is still sufficient. This is typically a problem one should take into consideration when designing the filter. With the finite window of length N , the recursion (24.94) essentially reduces to a *finite-memory* filter.

Instead of using a finite window, one could also consider using a fading window [24.93, 99, 100]. By setting $l = 1$ and replacing the gain g_k by

$$\frac{\omega^k}{\sum_{i=1}^k m_i \omega^i} ,$$

with weight $\omega > 1$, the recursion reduces to a *fading-memory* filter. Note that with the fading window we still have $E(T_{\text{GOM}}^{l,k} | H_0) = 1$. Instead of a central F -distribution, the fading GOM-test statistic will now follow a linear combination of independent χ^2 -distributions. With the fading window, the same recursion (24.94) is retained. This becomes advantageous when compared to the finite window if a particular application requires the use of long windows. The weight ω , which determines the nominal length of the fading window, is chosen on the basis of the detection power of the GOM test.

24.5.5 Recursive Identification

Local Identification

Recall that the local alternative is of the form (24.89)

$$\mathcal{H}_a^k : E(\mathbf{v}_k) = \mathbf{C}_{v_k} \mathbf{b} . \quad (24.95)$$

The type of model error considered is determined by the choice of the $m_k \times q$ matrix \mathbf{C}_{v_k} . Any model error that biases \mathbf{v}_k can be considered. These could be biases in the vector of observables \mathbf{y}_k and/or in the predicted state

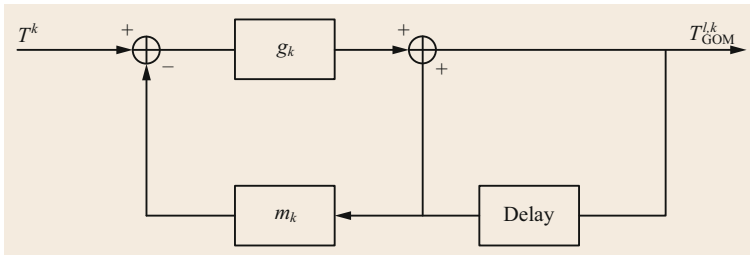


Fig. 24.17 The recursion of the GOM test statistic

$\hat{\mathbf{x}}_{k|k-1}$. For an outlier in the i th entry of \mathbf{y}_k , for instance, the matrix \mathbf{C}_{v_k} reduces to the canonical unit vector having a one as its i th entry.

All test statistics for (24.95) follow from (24.88) by setting $\mathbf{C}_{v_i} = \mathbf{0}$ for $i < k$. For the case $q = 1$, for example for data snooping at epoch k , one can make use of the normally distributed statistic

$$t^k = \frac{\mathbf{c}_{v_k}^T \mathbf{Q}_{v_k}^{-1} \mathbf{v}_k}{\sqrt{\mathbf{c}_{v_k}^T \mathbf{Q}_{v_k}^{-1} \mathbf{c}_{v_k}}} \overset{\mathcal{H}_0}{\sim} \mathcal{N}(0, 1) \quad (24.96)$$

since $T_{q=1}^k = (t^k)^2$.

Global Identification

The global alternative is of the form (24.90)

$$\mathcal{H}_a^{l,k} : E(\mathbf{v}^{l,k}) = \mathbf{C}_v^{l,k} \mathbf{b}. \quad (24.97)$$

All test statistics for this global alternative follow from (24.88) by setting $\mathbf{C}_{v_i} = \mathbf{0}$ for $i < l$. As they will be expressed in

$$\mathbf{C}_v^{l,k} = (\mathbf{C}_{v_l}^T, \dots, \mathbf{C}_{v_k}^T)^T,$$

we need a recursion for computing the entries of $\mathbf{C}_v^{l,k}$ so as to be able to perform the testing recursively. The recursion of \mathbf{C}_{v_i} , $i = l, \dots, k$, can be determined from the way the assumed bias propagates itself through the time and measurement updates of the Kalman filter. For instance, if it is assumed that the means of the observables are biased in the interval $[l, k]$, then we have under the alternative

$$\mathcal{H}_a^{l,k} : E(\mathbf{y}_i) = \mathbf{A}_i \mathbf{x}_i + \mathbf{C}_i \mathbf{b}, \quad i \in [l, k]. \quad (24.98)$$

As the bias $\mathbf{C}_i \mathbf{b}$ propagates through the Kalman filter, it biases the predicted residuals and the predicted state as $\mathbf{C}_{v_i} \mathbf{b}$ and $\hat{\mathbf{x}}_{i|i-1} \mathbf{b}$ respectively, with the response matrices satisfying the recursion (Fig. 24.18)

$$\begin{cases} \mathbf{C}_{v_i} = \mathbf{C}_i - \mathbf{A}_i \mathbf{C}_{\hat{\mathbf{x}}_{i|i-1}}, & \mathbf{C}_{\hat{\mathbf{x}}_{l|l-1}} = \mathbf{0}, \\ \mathbf{C}_{\hat{\mathbf{x}}_{i+1|i}} = \Phi_{i+1,i} (\mathbf{C}_{\hat{\mathbf{x}}_{i|i-1}} + \mathbf{K}_i \mathbf{C}_{v_i}), & i \geq l. \end{cases} \quad (24.99)$$

A likewise recursion can be found if, instead of the alternative (24.98), biases in the system noise are assumed present, for example due to an underparametrization of the dynamic model; see for example [24.25, 94].

As in the case of detection, the global identification test statistic can also be computed in recursive form. For $q = 1$, we have

$$(t^{l,k})^2 = (t^{l,k-1})^2 + h_{l,k} \left[(t^k)^2 - (t^{l,k-1})^2 \right], \quad (24.100)$$

with gain

$$h_{l,k} = \frac{\mathbf{c}_{v_k}^T \mathbf{Q}_{v_k}^{-1} \mathbf{c}_{v_k}}{\sum_{i=l}^k \mathbf{c}_{v_i}^T \mathbf{Q}_{v_i}^{-1} \mathbf{c}_{v_i}}.$$

Note that the local statistic t^k of (24.96) is used as input for each new epoch.

Strictly speaking the above test statistic has to be computed for each alternative hypothesis considered and for each epoch $k > l$. However, since l is unknown, one has to start in principle with $l = 1$. This implies that one has to compute k number of test statistics per alternative hypothesis at the time of testing k . As a result one obtains a test matrix of increasing order with $t^{l,k}$ as its entries. An example is shown in Fig. 24.19.

Such an increase in the number of test statistics is clearly unpractical, both from a computational point of view and because of the potential delay in time of identification. Fortunately not all entries of the test matrix may be necessary if one studies the power of the test statistics. Although the power will increase for longer intervals $[l, k]$, at a certain stage the gain in power may become negligible for all practical purposes. This motivates, in accordance with our discussion of detection, the use of a moving window. This is shown in Fig. 24.19b for the case $l \in [k-N+1, k]$, with $N = 2$, and in Fig. 24.19c for the case $l \in [k-N+1, k-M]$, with $N = 2$ and $M = 1$. The rationale behind this last constraint is that in some applications the test statistic may be too insensitive for global identification if $l > k-M$.

Once the test matrix has been defined, the identification procedure can proceed as follows. At the time of testing k , one first determines per alternative hypothesis the value of l in the window for which $|t^{l,k}|$ is at its maximum. Hence, the k th column of the test matrix is searched for the largest entry in absolute value. The corresponding matrix row number then identifies l as the most likely time of occurrence of the model error if the corresponding alternative hypothesis would be true. In order to find both the most likely alternative hypothesis and most likely value of l , the values of $\max_{l \in [k-N+1, k-M]} |t^{l,k}|$ for the different alternative hypotheses are compared. The maximum of this set finally identifies the most likely time of occurrence l and the most likely alternative hypothesis. Its likelihood is then tested against the chosen critical value or p -value.

24.5.6 Recursive Adaptation: General Case

After identification of the most likely model error, adaptation of the recursive filter is needed so as to reduce the presence of biases in the filtered state. If we assume that (24.97) has been identified as the most likely hypothesis, then – in analogy with (24.65) – the adaptation step

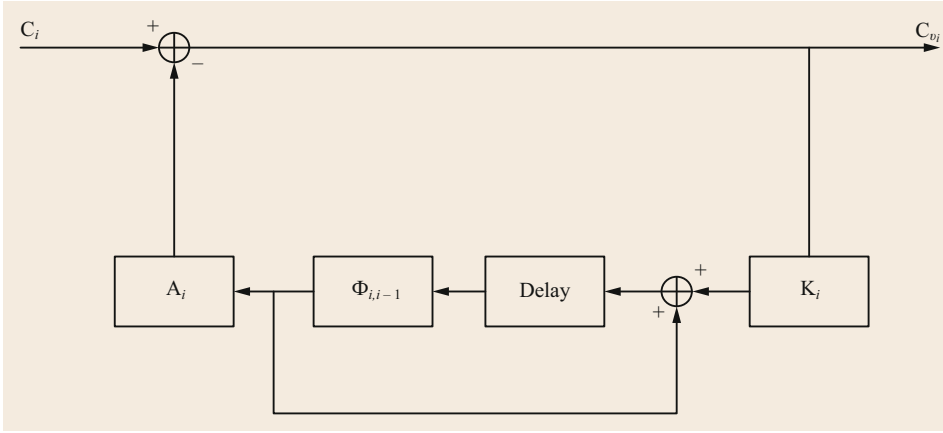


Fig. 24.18
Recursion of C_{v_i} in response to observation biases $C_i b$, with $i \in [l, k]$

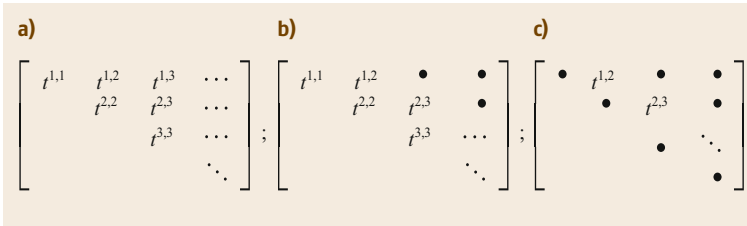


Fig. 24.19a–c Test matrices $t^{l,k}$ ($l \leq k$) for identification, with (a) no window, (b) moving window $l \in [k-1, k]$, and (c) moving window $l \in [k-1, k-1]$

is given as

$$\begin{aligned} \hat{x}_{k|k}^a &= \hat{x}_{k|k} - C_{\hat{x}_{k|k}} \hat{b}_{l|k}, \\ P_{k|k}^a &= P_{k|k} + C_{\hat{x}_{k|k}} Q_{l|k} C_{\hat{x}_{k|k}}^T, \end{aligned} \quad (24.101)$$

with $\hat{b}_{l|k}$ the least-squares solution of (24.97), $Q_{l|k}$ its variance matrix, and $C_{\hat{x}_{k|k}}$ the bias-response matrix of the filter state. By again making use of the block diagonal structure of the predicted residual vector's variance matrix, the bias estimator can be formulated in recursive form as

$$\begin{aligned} \hat{b}_{l|k} &= \hat{b}_{l|k-1} + G_k (v_k - C_{v_k} \hat{b}_{l|k-1}), \\ Q_{l|k} &= (I_q - G_k C_{v_k}) Q_{l|k-1}, \end{aligned} \quad (24.102)$$

with gain matrix

$$G_k = Q_{l|k-1} C_{v_k}^T (Q_{v_k} v_k + C_{v_k} Q_{l|k-1} C_{v_k}^T)^{-1}. \quad (24.103)$$

Hence, since both the bias estimator $\hat{b}_{l|k}$ and the filtered response matrix $C_{\hat{x}_{k|k}} = \Phi_{k,k+1} C_{\hat{x}_{k+1|k}}$ (cf. (24.99)) can be computed recursively, the adaptation (24.101) can be executed recursively as well. The recursion of the bias estimator and filtered response matrix is shown in Fig. 24.20.

The adaptation (24.101) reduces to the adaptation for (24.95) in the case where $l = k$. Thus if local iden-

tification can be done successfully, it has the advantage that immediate corrective action is possible by adapting the filter at the same time that the model error occurred. This is not possible in the case of global identification. As one is then confronted with a delay in time of identification (i. e., identifying at epoch k of an error that occurred at a previous epoch $l < k$), the filtered state will remain biased within the interval $[l, k]$. Whether this is considered acceptable depends on the application at hand. Correcting these filtered states would involve smoothing, which can be done recursively, but which may not be needed in real-time applications as it is a corrective action after the fact.

As adaptation implies that the identified alternative hypothesis takes over the role of the null hypothesis, the filtering from time k onwards will be based on the extended state vector and variance matrix

$$\begin{bmatrix} \hat{x}_{k|k}^a \\ \hat{b}_{l|k} \end{bmatrix}, \begin{bmatrix} P_{k|k}^a & -C_{\hat{x}_{k|k}} Q_{l|k} \\ -Q_{l|k} C_{\hat{x}_{k|k}}^T & Q_{l|k} \end{bmatrix} \quad (24.104)$$

the recursion of which is given by (24.101) and (24.102) respectively [24.75, 101].

24.5.7 Recursive Adaptation: Special GNSS Case

Not in all cases does the filtering after the adaptation step need to continue with an extended state vector as

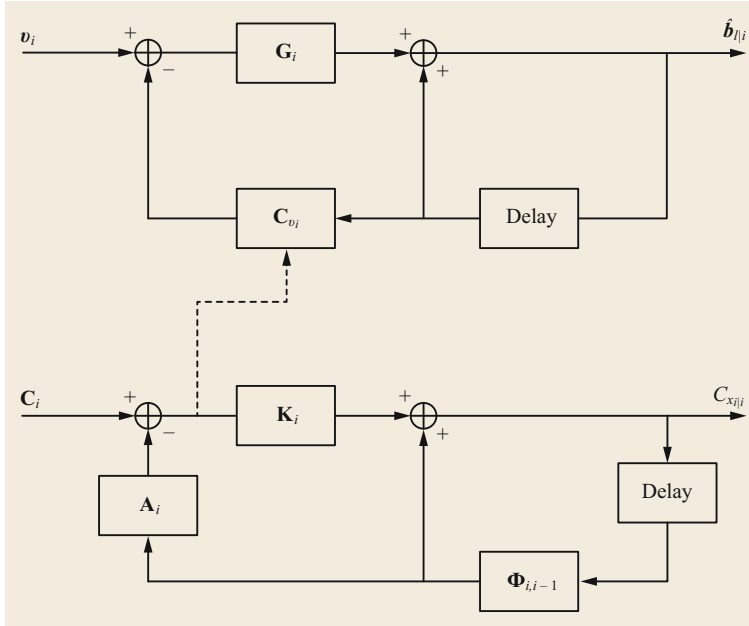


Fig. 24.20 Recursion of the bias estimator $\hat{\mathbf{b}}_{l|i}$ and filtered response matrix $\mathbf{C}_{\hat{\mathbf{x}}_{l|i}}$

in (24.104). There are two important GNSS biases for which this is not needed. They are the biases due to code outliers and the biases due to carrier-phase slips. In both cases the adaptation needs to be performed only once, after which one can revert back again to the standard filtering under the null hypothesis.

Code Outliers

Consider the case that the assumed modeling error in the observables is confined to a single epoch l . The \mathbf{C}_i -matrix of (24.98) is then given as

$$\mathbf{C}_i = \mathbf{C}_l \delta_{i,l}, \quad i = 1, \dots, k. \quad (24.105)$$

A special case of this structure is the occurrence of a code outlier, for which matrix \mathbf{C}_l becomes the canonical unit vector having its nonzero entry at the place of the suspected code observable.

As spike-like modeling errors of the type (24.105) are not persistent, but instead confined to a single epoch, their adaptation only needs to be done once. After such adaptation one can then simply proceed again with the filtering under \mathcal{H}_0 . The adapted state and its (error) variance matrix, $\hat{\mathbf{x}}_{k|k}^a$ and $\mathbf{P}_{k|k}^a$, are then in fact treated as a new initialization of the standard filter.

Carrier-Phase Slips

For a slip, the \mathbf{C}_i -matrix of (24.98) is given as

$$\mathbf{C}_i = \mathbf{C}_l s_{i,l}, \quad i = 1, \dots, k, \quad (24.106)$$

with the step function $s_{i,l} = 0$ for $i < l$ and $s_{i,l} = 1$ for $i \geq l$. Since slips are persistent, one generally will have to work with the extended state vector (24.104) for all following epochs. However, an important exception exists for GNSS carrier-phase slips, where filtering under \mathcal{H}_0 is possible after a single adaptation step. This property is a consequence of the fact that one can show that reverting back to filtering under \mathcal{H}_0 is possible for any slip that can be parametrized as

$$\mathbf{C}_i = \mathbf{A}_i \boldsymbol{\Phi}_{i,l} \mathbf{X}_l s_{i,l} \text{ for some } \mathbf{X}_l \in \mathbb{R}^{n \times q}. \quad (24.107)$$

By then reparametrizing the state vector under \mathcal{H}_a as

$$\begin{bmatrix} \bar{\mathbf{x}}_i \\ \bar{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \boldsymbol{\Phi}_{i,l} \mathbf{X}_l \\ \mathbf{0} & \mathbf{I}_q \end{bmatrix} \begin{bmatrix} \mathbf{x}_i \\ \mathbf{b} \end{bmatrix} \quad (24.108)$$

the adaptation at epoch k takes the form

$$\begin{aligned} \hat{\mathbf{x}}_{k|k}^a &= \hat{\mathbf{x}}_{k|k} - \bar{\mathbf{C}}_{\hat{\mathbf{x}}_{k|k}} \hat{\mathbf{b}}_{l|k}, \\ \bar{\mathbf{P}}_{k|k}^a &= \mathbf{P}_{k|k} + \bar{\mathbf{C}}_{\hat{\mathbf{x}}_{k|k}} \mathbf{Q}_{l|k} \bar{\mathbf{C}}_{\hat{\mathbf{x}}_{k|k}}^T, \end{aligned} \quad (24.109)$$

with $\bar{\mathbf{C}}_{\hat{\mathbf{x}}_{k|k}} = (\mathbf{C}_{\hat{\mathbf{x}}_{k|k}} - \boldsymbol{\Phi}_{k,l} \mathbf{X}_l)$. Compare with (24.101).

Since the reparametrization (24.108) compensates rigorously for the bias effect of \mathbf{b} , the adaptation step (24.109) only needs to be done once, after which one can continue with the filtering under \mathcal{H}_0 . Thus with the new initialization, $\hat{\mathbf{x}}_{k|k}^a$ and $\bar{\mathbf{P}}_{k|k}^a$, one can continue with the standard filter under \mathcal{H}_0 , thereby recognizing

of course that the state vector \mathbf{x}_i has now been replaced by its reparametrized version $\bar{\mathbf{x}}_i$ for $i \geq k$.

The above results directly apply to the case of GNSS carrier-phase slips. To see this, let $\mathbf{y}_i = [\mathbf{p}_i^T, \boldsymbol{\phi}_i^T]^T$ be the code (\mathbf{p}_i) and phase ($\boldsymbol{\phi}_i$ in cycles) observation vector at epoch i , with partitioned design matrix and state vector, $\mathbf{A}_i = [\mathbf{A}_{1i}, \mathbf{A}_{2i}]$ and $\mathbf{x}_i = [\mathbf{x}_{1i}^T, \mathbf{x}_{2i}^T]^T$ respectively, in which \mathbf{x}_2 is the time-constant ambiguity vector in cycles. Then $\mathbf{A}_{2i} = [\mathbf{0}, \mathbf{I}_p]^T$ and $\boldsymbol{\phi}_{i,l} = \text{blockdiag}[\boldsymbol{\phi}_{1i,l}, \mathbf{I}_p]$. With a cycle slip at epoch l in the j th phase observable, we have $\mathbf{C}_i = [\mathbf{0}, \mathbf{c}_j^T]^T s_{i,l}$, thus showing that (24.107) is satisfied for $\mathbf{X}_l = [\mathbf{0}, \mathbf{c}_j^T]^T$. With $\boldsymbol{\phi}_{i,l} \mathbf{X}_l = [\mathbf{0}, \mathbf{c}_j^T]^T$, the reparametrization (24.108) for $\bar{\mathbf{x}}_i$ then simplifies to

$$\begin{bmatrix} \bar{\mathbf{x}}_{1i} \\ \bar{\mathbf{x}}_{2i} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{c}_j \end{bmatrix} b, \quad (24.110)$$

thus showing that in this case it only affects the ambiguities and not the other state vector entries.

With $\bar{\mathbf{C}}_{\hat{\mathbf{x}}_{k|k}} = [\mathbf{C}_{\hat{\mathbf{x}}_{k|k}} - [\mathbf{0}, \mathbf{c}_j^T]^T]$, the carrier-phase slip adaptation then takes the simple form

$$\begin{bmatrix} \hat{\mathbf{x}}_{1k|k} \\ \hat{\mathbf{x}}_{2k|k} \end{bmatrix}^a = \begin{bmatrix} \hat{\mathbf{x}}_{1k|k} \\ \hat{\mathbf{x}}_{2k|k} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_{1\hat{\mathbf{x}}_{k|k}} \\ \mathbf{C}_{2\hat{\mathbf{x}}_{k|k}} - \mathbf{c}_j \end{bmatrix} \hat{b}_{l|k}, \quad (24.111)$$

in which $\hat{b}_{l|k}$ is the estimated slip. Thus when after adaptation one reverts back to the standard filter under \mathcal{H}_0 , the interpretation of all state vector entries remains the same except for the ambiguities. The adapted ambiguities differ from the original ambiguities by the slip.

Acknowledgments. The author is the recipient of an Australian Research Council Federation Fellowship (project number FF0883188). Ms Safoora Zaminpardaz and Dr Amir Khodabandeh of Curtin's GNSS Research Centre helped with the examples and the creation of figures. All this support is gratefully acknowledged.

References

- 24.1 G. Blewitt: An automatic editing algorithm for GPS data, *Geophys. Res. Lett.* **17**(3), 199–202 (1990)
- 24.2 N. Zinn, P.J.V. Rapatz: Reliability analysis in marine seismic networks, *Hydrogr. J.* **76**, 11–18 (1995)
- 24.3 R.G. Brown: Receiver autonomous integrity monitoring. In: *Global Positioning System: Theory and applications*, Vol. 2, ed. by B.W. Parkinson, J.J. Spilker Jr. (American Institute of Aeronautics and Astronautics, Washington 1996) pp. 143–165
- 24.4 P.J.G. Teunissen: Quality control and GPS. In: *GPS for Geodesy*, ed. by P.J.G. Teunissen, A. Kleusberg (Springer, Berlin Heidelberg 1998) pp. 271–318
- 24.5 A. Leick: *GPS Satellite Surveying*, Vol. 3 (Wiley, New Jersey 2004)
- 24.6 A. Wieser, M.G. Petovello, G. Lachapelle: Failure scenarios to be considered with kinematic high precision relative GNSS positioning, *Proc. ION GNSS 2004*, Long Beach (ION, Virginia 2004) pp. 1448–1459
- 24.7 B. Hofmann-Wellenhof, H. Lichtenegger, E. Wasle: *GNSS—Global Navigation Satellite Systems: GPS, GLONASS, Galileo, and More* (Springer, New York 2008)
- 24.8 J. Oliveira, C.C.J.M. Tiberius: Quality control in SBAS: protection levels and reliability levels, *J. Navig.* **62**(3), 509–522 (2009)
- 24.9 S. Banville, R.B. Langley: Instantaneous Cycle-Slip Correction for Real-Time PPP Applications, *Navigation* **57**(4), 325–334 (2010)
- 24.10 C.D. De Jong: A unified approach to real-time integrity monitoring of single- and dual-frequency GPS and GLONASS observations, *Acta Geod. Geophys. Hung.* **33**(2–4), 247–257 (1998)
- 24.11 N.F. Jonkman, K. De Jong: Integrity monitoring of IIGX-98 data, Part II: Cycle slip and outlier detection, *GPS Solutions* **3**(4), 24–34 (2000)
- 24.12 C.D. De Jong, H. Van Der Marel, N.F. Jonkman: Real-time GPS and GLONASS integrity monitoring and reference station software, *Phys. Chem. Earth A* **26**(6), 545–549 (2001)
- 24.13 P.J.G. Teunissen, P.F. De Bakker: Single-receiver single-channel multi-frequency GNSS integrity: Outliers, slips, and ionospheric disturbances, *J. Geod.* **87**(2), 161–177 (2013)
- 24.14 Z.F. Biacs, E.J. Krakiwsky, D. Lapucha: Reliability analysis of phase observations in GPS baseline estimation, *J. Surv. Eng.* **116**(4), 204–224 (1990)
- 24.15 R.G. Brown: A baseline GPS RAIM scheme and a note on the equivalence of three RAIM methods, *Navigation* **39**(3), 301–316 (1992)
- 24.16 P.A. Cross, D.J. Hawksbee, R. Nicolai: Quality measures for differential GPS positioning, *Hydrogr. J.* **72**, 17–22 (1994)
- 24.17 C.D. De Jong: Real-time integrity monitoring of dual-frequency GPS observations for a single receiver, *Acta Geod. Geophys. Hung.* **31**(1/2), 37–46 (1996)
- 24.18 A. Wieser: Reliability checking for GNSS baseline and network processing, *GPS Solut.* **8**(2), 55–66 (2004)
- 24.19 H. Kuusniemi, A. Wieser, G. Lachapelle, J. Takala: User-level reliability monitoring in urban personal satellite-navigation, *IEEE Trans. Aerosp. Electron. Syst.* **43**(4), 1305–1318 (2007)
- 24.20 H. Van der Marel, A.J.M. Kusters: Statistical testing and quality analysis in 3-D networks: (Part II) Application to GPS. In: *Global Positioning System: An Overview*, ed. by Y. Bock, N. Leppard (Springer, New York 1990) pp. 290–297

- 24.21 K. De Jong: A modular approach to precise GPS positioning, *GPS Solutions* **2**(4), 52–56 (1999)
- 24.22 N. Perfetti: Detection of station coordinate discontinuities within the Italian GPS fiducial network, *J. Geod.* **80**(7), 381–396 (2006)
- 24.23 B.W. Parkinson, J.J. Spilker: *Global Positioning System: Theory and Applications* (AIAA, Washington 1996)
- 24.24 P.J.G. Teunissen, A. Kleusberg: *GPS for Geodesy*, Vol. 2 (Springer, Berlin, Heidelberg 1998)
- 24.25 M.A. Salzmann: Least squares filtering and testing for geodetic navigation applications, Ph.D. Thesis (Delft University of Technology, Delft 1993)
- 24.26 V. Gikas, P.A. Cross, D. Ridyard: Reliability analysis in dynamic systems: Implications for positioning marine seismic networks, *Geophysics* **64**(4), 1014–1022 (1999)
- 24.27 R.J. Kelly: Comparison of LAAS B-values with linear model optimum B-values, *Navigation* **47**(2), 143–156 (2000)
- 24.28 T. Murphy, M. Harris, C. Shively, L. Azoulai, M. Brenner: Fault modeling for GBAS airworthiness assessments, *Navigation* **59**(2), 145–161 (2012)
- 24.29 P.J.G. Teunissen: *Testing Theory: An Introduction* (Delft University Press, Delft 2000)
- 24.30 S.F. Arnold: *The Theory of Linear Models and Multivariate Analysis*, Vol. 2 (Wiley, New York 1981)
- 24.31 K.R. Koch: *Parameter Estimation and Hypothesis Testing in Linear Models* (Springer, Berlin 1999)
- 24.32 J. Neyman, E.S. Pearson: *On the Problem of the Most Efficient Tests of Statistical Hypotheses* (Springer, New York 1992)
- 24.33 W. Baarda: Statistical Concepts in Geodesy, Publications on Geodesy **2**(4) (Netherlands Geodetic Commission, Delft 1967)
- 24.34 W. Baarda: A testing procedure for use in geodetic networks, Publications on Geodesy **2**(5) (Netherlands Geodetic Commission, Delft 1968)
- 24.35 P.J.G. Teunissen: Quality control in integrated navigation systems, *IEEE Aerosp. Electron. Syst. Mag.* **5**(7), 35–41 (1989)
- 24.36 Staff of DGCC: The Delft Approach for the Design and Computation of Geodetic Networks, “Forty Years of Thought...” Anniversary edition on the occasion of the 65th birthday of Professor W. Baarda **1** (Delft University of Technology 1982) pp. 202–274
- 24.37 J. Van Mierlo: A testing procedure for analytic deformation measurements, *Proc. 2nd Int. Symp. Deform. Meas. Geod. Methods*, Bonn, ed. by L. Hallermann (Verlag Konrad Wittwer, Stuttgart 1981) pp. 321–353
- 24.38 J.J. Kok: Statistical analysis of deformation problems using Baarda’s testing procedures. In: “Forty Years of Thought”. Anniversary Volume on the Occasion of Prof. Baarda’s 65th Birthday, Delft **2** (Delft University of Technology 1982) pp. 470–488
- 24.39 W. Forstner: Reliability and discernability of extended Gauss–Markov models, *Semin. Math. Models Geod. Photogramm. Point Determ. Regard Outliers Syst. Errors*, Stuttgart, ed. by F.E. Ackermann (Deutsche Geodätische Kommission, München 1983) pp. 79–104
- 24.40 P.J.G. Teunissen: Adjusting and testing with the models of the affine and similarity transformation, *Manuscr. Geod.* **11**, 214–225 (1986)
- 24.41 G. Lu, G. Lachapelle: Reliability Analysis Applied to Kinematic GPS Position and Velocity Estimation, *Proc. Int. Symp. Kinemat. Syst. Geod. Surv. Remote Sens.* (KIS 1991), Banff, ed. by K.–P. Schwarz, G. Lachapelle (Springer, New York 1991) pp. 273–284
- 24.42 P.J.G. Teunissen: Minimal detectable biases of GPS data, *J. Geod.* **72**(4), 236–244 (1998)
- 24.43 K. De Jong: Minimal detectable biases of cross-correlated GPS observations, *GPS Solutions* **3**(3), 12–18 (2000)
- 24.44 K. O’Keefe, S. Ryan, G. Lachapelle: Global availability and reliability assessment of the GPS and Galileo global navigation satellite systems, *Can. Aeronaut. Space J.* **48**(2), 123–132 (2002)
- 24.45 K. De Jong, P.J.G. Teunissen: Minimal Detectable Biases of GPS observations for a weighted ionosphere, *Earth Planets Space* **52**(10), 857–862 (2000)
- 24.46 S. Verhagen: Performance analysis of GPS, Galileo and integrated GPS–Galileo, *Proc. ION GPS 2002*, Portland (ION, Virginia 2002) pp. 2208–2215
- 24.47 F. Wu, N. Kubo, A. Yasuda: Performance analysis of GPS augmentation using Japanese quasi-zenith satellite system, *Earth Planets Space* **56**(1), 25–37 (2004)
- 24.48 S. Hewitson, H. Kyu Lee, J. Wang: Localizability analysis for GPS/Galileo receiver autonomous integrity monitoring, *J. Navig.* **57**(02), 245–259 (2004)
- 24.49 C. Zhao, J. Ou, Y. Yuan: Positioning accuracy and reliability of GALILEO, integrated GPS–GALILEO system based on single positioning model, *Chin. Sci. Bull.* **50**(12), 1252–1260 (2005)
- 24.50 S. Hewitson, J. Wang: GNSS receiver autonomous integrity monitoring (RAIM) performance analysis, *GPS Solutions* **10**(3), 155–170 (2006)
- 24.51 H. Xu, J. Wang, X. Zhan: GNSS Satellite Autonomous Integrity Monitoring (SAIM) using intersatellite measurements, *Adv. Space Res.* **47**(7), 1116–1126 (2011)
- 24.52 X. Su, X. Zhan, M. Niu, Y. Zhang: Receiver autonomous integrity monitoring availability and fault detection capability comparison between BeiDou and GPS, *J. Shanghai Jiaotong Univ. (Sci.)* **19**(3), 313–324 (2014)
- 24.53 M.A. Salzmann: MDB: A design tool for integrated navigation systems, *Bull. Geod.* **65**(2), 109–115 (1991)
- 24.54 P.J.G. Teunissen: Internal reliability of single frequency GPS data, *Artif. Satell.* **32**(2), 63–73 (1997)
- 24.55 J.E. Alberda: Quality control in surveying, *Chart. Surv.* **4**(2), 23–28 (1976)
- 24.56 P.J.G. Teunissen: Quality control in geodetic networks. In: *Optimization and Design of Geodetic Networks*, ed. by E. Grafarend, F. Sanso (Springer, Berlin 1985) pp. 526–547

- 24.57 C.C.J.M. Tiberius: *Recursive Data Processing for Kinematic GPS Surveying*, Publication on Geodesy, Vol. 45 (Nederlandse Commissie Voor Geodesie, Delft 1998)
- 24.58 P.B. Ober: *Integrity Prediction and Monitoring of Navigation Systems*, Vol. 1 (Integricom Publishers, Leiden 2003)
- 24.59 B. Kargoll: On the theory and application of model misspecification tests in geodesy, Ph.D. Thesis (Rheinische Friedrichs-Wilhelms-Universität, Bonn 2007)
- 24.60 C. Robert, G. Casella: *Monte Carlo Statistical Methods* (Springer Science and Business Media, New York 2013)
- 24.61 H. Scheffé: *The Analysis of Variance* (Wiley, New York 1999)
- 24.62 R.G. Miller: *Simultaneous Statistical Inference*, 2nd edn. (Springer, New York, Heidelberg Berlin 1981)
- 24.63 P.H. Westfall: *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, Vol. 279 (Wiley, New York 1993)
- 24.64 A.J. Pope: The statistics of residuals and the detection of outliers, NOAA Technical Report NOS 65 NGS 1 (US Department of Commerce, NOAA, Rockville 1976)
- 24.65 D.M. Hawkins: *Identification of outliers*, Vol. 11 (Chapman and Hall, London 1980)
- 24.66 R.J. Beckman, R.D. Cook: Outlier s, *Technometrics* **25**(2), 119–149 (1983)
- 24.67 B.W. Parkinson, P. Axelrad: Autonomous GPS integrity monitoring using the pseudorange residual, *Navigation* **35**(2), 255–274 (1988)
- 24.68 M.A. Sturza: Navigation system integrity monitoring using redundant measurements, *Navigation* **35**(4), 483–501 (1988)
- 24.69 P.J.G. Teunissen: Differential GPS: Concepts and Quality Control, NIN Workshop Navstar GPS, Amsterdam (Delft Geodetic Computing Centre LGR, Delft 1991) pp. 1–49
- 24.70 V. Barnett, T. Lewis: *Outliers in Statistical Data*, Vol. 3 (Wiley, New York 1994)
- 24.71 R. J. Kelly: The linear model, RNP, and the near-optimum fault detection and exclusion algorithm, *Glob. Position. Syst. ION Red Book Series*, Vol. 5, (ION, Manassas 1998) pp. 227–259
- 24.72 H. Kuusniemi, G. Lachapelle, J.H. Takala: Position and velocity reliability testing in degraded GPS signal environments, *GPS Solutions* **8**(4), 226–237 (2004)
- 24.73 M. Kern, T. Preimesberger, M. Allesch, R. Pail, J. Bouman, R. Koop: Outlier detection algorithms and their performance in GOCE gravity field processing, *J. Geod.* **78**(9), 509–519 (2005)
- 24.74 J.J. Kok: On data snooping and multiple outlier testing, NOAA Technical Report NOS NGS 30 (US Department of Commerce, NOAA, Rockville 1984)
- 24.75 P.J.G. Teunissen: An integrity and quality control procedure for use in multi sensor integration, *Proc. ION GPS 1990*, Colorado Springs (ION, Virginia 1990) pp. 513–522
- 24.76 X. Ding, R. Coleman: Multiple outlier detection by evaluating redundancy contributions of observations, *J. Geod.* **70**(8), 489–498 (1996)
- 24.77 B.S. Pervan, S.P. Pullen, J.R. Christie: A multiple hypothesis approach to satellite navigation integrity, *Navigation* **45**(1), 61–71 (1998)
- 24.78 J.E. Angus: RAIM with multiple faults, *Navigation* **53**(4), 249–257 (2006)
- 24.79 S. Hewitson, J. Wang: GNSS receiver autonomous integrity monitoring (RAIM) for multiple outliers, *Eur. J. Navig.* **4**(4), 47–57 (2006)
- 24.80 J. Blanch, T. Walter, P. Enge: RAIM with optimal integrity and continuity allocations under multiple failures, *IEEE Trans. Aerosp. Electron. Syst.* **46**(3), 1235–1247 (2010)
- 24.81 D. Imparato: Detecting multi-dimensional threats: A comparison of solution separation test and uniformly most powerful invariant test, *Proc. Eur. Navig. Conf. (ENC)-GNSS 2014*, Rotterdam (Nederlands Instituut voor Navigatie, Nederlands 2014) pp. 1–13
- 24.82 D. Imparato: GNSS Based Receiver Autonomous Integrity Monitoring for Aircraft Navigation, Ph.D. Thesis (TU Delft, Delft 2016)
- 24.83 C.E. Bonferroni: Teoria statistica delle classi e calcolo delle probabilit , *Pubblicazioni del R Istituto Super. di Scienze Econ. e Commer. di Firenze* **8**, 3–62 (1936)
- 24.84 Y. Benjamini, Y. Hochberg: Controlling the false discovery rate: A practical and powerful approach to multiple testing, *J. R. Stat. Soc. B* **57**(1), 289–300 (1995)
- 24.85 Y. Benjamini, D. Yekutieli: The control of the false discovery rate in multiple testing under dependency, *Ann. Stat.* **29**(4), 1165–1188 (2001)
- 24.86 B. Efron: Correlation and large-scale simultaneous significance testing, *J. Am. Stat. Assoc.* **102**(477), 93–103 (2007)
- 24.87 P.B. Ober: Integrity according to Bayes, *Proc. IEEE PLANS 2000*, San Diego (IEEE, Piscataway 2000) pp. 325–332, doi:10.1109/PLANS.2000.838321
- 24.88 C.R. Rao: *Linear Statistical Inference and Its Applications* (Wiley, New York 1973)
- 24.89 W. Gosset: (Student): The probable error of a mean, *Biometrika* **6**(1), 1–25 (1908)
- 24.90 R.E. Kalman: A new approach to linear filtering and prediction problems, *J. Basic Eng.* **82**(1), 35–45 (1960)
- 24.91 A. Gelb: *Applied Optimal Estimation* (MIT Press, Cambridge 1974)
- 24.92 B.D.O. Anderson, J.B. Moore: *Optimal Filtering* (Prentice-Hall, Englewood Cliffs, New Jersey 1979)
- 24.93 A.H. Jazwinski: *Stochastic Processes and Filtering Theory* (Dover Publications, New York 1991)
- 24.94 P.J.G. Teunissen, M.A. Salzmann: A recursive slip-page test for use in state-space filtering, *Manuscr. Geod.* **14**(6), 383–390 (1989)
- 24.95 I. Gillissen, I.A. Elema: Test results of DIA: A real-time adaptive integrity monitoring procedure, used in an integrated navigation system, *Int. Hydrogr. Rev.* **73**(1), 75–103 (1996)

- 24.96 G. Lu, G. Lachapelle: Statistical quality control for kinematic GPS positioning, *Manuscr. Geod.* **17**(5), 270–281 (1992)
- 24.97 S. Hewitson, J. Wang: GNSS receiver autonomous integrity monitoring with a dynamic model, *J. Navig.* **60**(02), 247–263 (2007)
- 24.98 J.G. Wang: Reliability analysis in Kalman filtering, *J. Glob. Position. Syst.* **8**(1), 101–111 (2009)
- 24.99 H.W. Sorenson, J.E. Sacks: Recursive fading memory filtering, *Inf. Sci.* **3**(2), 101–119 (1971)
- 24.100 B.D.O. Anderson: Exponential Data Weighting in the Kalman–Bucy Filter, *Inf. Sci.* **5**, 217–230 (1973)
- 24.101 M.A. Salzmann: Real-time adaptation for model errors in dynamic systems, *Bull. Geod.* **69**(2), 81–91 (1995)

Positioning

Part E

Part E Positioning and Navigation

25 Precise Point Positioning

Jan Kouba, Ottawa, Canada
François Lahaye, Ottawa, Canada
Pierre Tétreault, Ottawa, Canada

26 Differential Positioning

Dennis Odijk, Leidschendam,
The Netherlands
Lambert Wanninger, Dresden, Germany

27 Attitude Determination

Gabriele Giorgi, Munich, Germany

28 GNSS/INS Integration

Jay A. Farrell, Riverside, USA
Jan Wendel, Taufkirchen, Germany

29 Land and Maritime Applications

Allison Kealy, Parkville, Australia
Terry Moore, Nottingham, UK

30 Aviation Applications

Richard Farnworth, Bretigny sur Orge,
France

31 Ground Based Augmentation Systems

Sam Pullen, Stanford, USA

32 Space Applications

Oliver Montenbruck, Wessling, Germany

Precise Point Positioning

Jan Kouba, François Lahaye, Pierre Tétreault

Since its introduction in 1997, precise point positioning (PPP) offers an attractive alternative to differential global navigation satellite system (GNSS) positioning. The PPP approach uses undifferenced, dual-frequency, pseudorange and carrier-phase observations along with precise satellite orbit and clock products, for standalone static or kinematic geodetic point positioning with centimeter precision. This chapter introduces the PPP concept and specifies the required models needed to correct for systematic effects causing centimeter-level variations in the satellite-to-user range. For completeness, models and methods for processing single-frequency GNSS data are presented and specific aspects of GLONASS (Global'naya Navigatsionnaya Sputnikova Sistema) and new GNSSs are also described. Furthermore, recent developments in fixing undifferenced carrier-phase ambiguities, which can considerably shorten or nearly eliminate the initial delay for PPP convergence, are highlighted. Existing web applications and real-time corrections services enabling post-mission and real-time PPP are presented. Finally, typical PPP precision and accuracy estimates are discussed, including the solution of station tropospheric zenith path delays and receiver clocks, with millimeter and nanosecond precision respectively.

25.1 PPP Concept	724
25.1.1 Observation Equations	724
25.1.2 Adjustment and Quality Control	725
25.2 Precise Positioning Correction Models	726
25.2.1 Atmospheric Propagation Delays	728
25.2.2 Antenna Effects	730
25.2.3 Site Displacement Effects	732
25.2.4 Differential Code Biases	733
25.2.5 Compatibility and Conventions	734
25.3 Specific Processing Aspects	735
25.3.1 Single-Frequency Positioning	735
25.3.2 GLONASS PPP Considerations	736
25.3.3 New Signals and Constellations	737
25.3.4 Phase Ambiguity Fixing in PPP	739
25.4 Implementations	741
25.4.1 Post-Processed Solutions	741
25.4.2 Real-Time Solutions	742
25.4.3 PPP Positioning Services	742
25.5 Examples	743
25.5.1 Static PPP Solutions	743
25.5.2 Kinematic PPP Solutions	743
25.5.3 Tropospheric Zenith Path Delay	745
25.5.4 Station Clock Solutions	745
25.6 Discussion	746
References	747

The potential of GNSS for geodetic positioning applications was realized quite early during the Global Positioning System (GPS) implementation stage [25.1]. A relative positioning method, utilizing carrier-phase measurements made simultaneously and doubly differenced (DD) between two observing stations and two satellites, was proposed to eliminate the satellite and receiver clock offsets. Until the mid-1990s, practically all geodetic GPS applications employed relative baseline positioning with DD carrier-phase observations (Chap. 26).

In 1997, a new approach called precise point positioning (PPP), utilizing undifferenced carrier-phase

and pseudorange observations was introduced by Zumberge et al. [25.2]. Unlike the traditional DD relative baseline positioning, PPP does not require simultaneous observations at two stations. PPP, in fact, is a logical extension of the GNSS pseudorange navigation, which replaces the broadcast satellite orbits and clocks with precise estimates, and includes the precise carrier-phase observations in addition to the pseudoranges. This, however, necessitates the introduction of additional initial phase ambiguity unknowns, causing a fairly long (up to 15 min or longer) initial convergence of PPP solutions. It also entails careful modeling and data screening for outliers and carrier-phase cycle

slips, which is more challenging than for the DD approach.

PPP also requires much more careful modeling of local station and environmental effects than DD relative positioning. However, in addition to precise position solutions, PPP provides precise station clocks and tropospheric zenith path delays (ZTDs), which are either unavailable or less precise in the case of DD positioning. The greater availability of precise orbit and clock solution products in late 1990s, thanks in great part to the organized efforts of the International GNSS Service (IGS); (Chap. 33), increased the popularity of PPP for geodetic and many other applications, for example in geodynamics, meteorology, metrology [25.3] and so on. This is clearly demonstrated by the popularity of the

several online PPP services and PPP software packages now available.

The purpose of this chapter is to provide an overview of the PPP concept, state-of-the art PPP modeling techniques and the achievable performances. In Sect. 25.1 the PPP concept is introduced, followed by Sect. 25.2, which discusses conventional correction models and compatibility aspects. This is complemented by a review of specific processing aspects such as single-frequency and multi-GNSS PPP as well as the recent developments of precise point positioning using undifferenced phase ambiguity fixing (Sect. 25.3). The last two sections, 25.4 and 25.5, respectively list available PPP implementations and services and provide more detailed examples of recent PPP results.

25.1 PPP Concept

The PPP approach assumes that globally consistent satellite orbits and clocks are fixed or heavily constrained, and that PPP mathematical models are consistent with those applied in the global network solutions from which the orbit/clock products were estimated. In general, this consistency can be readily achieved if both the global orbit/clock and PPP solutions adhere to the same international standards, such as the current International Earth Rotation and Reference Systems Service (IERS) conventions. Since carrier-phase observations are used, PPP must estimate initial phase ambiguities to all satellites, in addition to the station position, station clock offsets and tropospheric zenith path delays (ZTD). The PPP method can be conceptualized as a back substitution of single station data into a global solution condensed in the form of the global satellite orbits and clocks and associated conventions and standards. Although PPP itself uses data from a single station only, computation of the satellite orbits and clocks needed for its implementation require the use of a global tracking network.

25.1.1 Observation Equations

For PPP, typically dual-frequency data is combined in order to eliminate nearly all of the ionospheric propagation delays. The ionosphere-free (IF) combinations (Chap. 19) of dual-frequency GNSS pseudorange (p_{IF}) and carrier-phase observations (φ_{IF}) are related to the user position, clock, troposphere and ambiguity parameters according to the following simplified observation equations (Chap. 19)

$$\begin{aligned} p_{\text{r,IF}}^s &= \rho_r^s + c (dt_r - dt^s) + T_r^s + e_{\text{IF}}^s, \\ \varphi_{\text{r,IF}}^s &= \rho_r^s + c (dt_r - dt^s) + T_r^s + \lambda_{\text{IF}} A_{\text{IF}} + \epsilon_{\text{IF}}^s, \end{aligned} \quad (25.1)$$

where:

- $p_{\text{r,IF}}^s$ is the ionosphere-free combination $(f_A^2 p_A - f_B^2 p_B) / (f_A^2 - f_B^2)$ of pseudoranges p_A and p_B observed at two distinct signal frequencies f_A and f_B .
- $\varphi_{\text{r,IF}}^s$ is the ionosphere-free combination $(f_A^2 \varphi_A - f_B^2 \varphi_B) / (f_A^2 - f_B^2)$ of the corresponding carrier-phases φ_A and φ_B .
- ρ_r^s is the geometrical range $\|\mathbf{x}^s - \mathbf{x}_r\|$ from the satellite position $\mathbf{x}^s = (x^s, y^s, z^s)^\top$ at the signal emission epoch t_E to the receiver position $\mathbf{x}_r = (x_r, y_r, z_r)^\top$ at its reception (arrival) epoch $t_A \cong t_E + \rho_r^s / c$.
- dt_r is the receiver clock offset from the GNSS time (including receiver code biases and delays).
- dt^s is the satellite clock offset from the GNSS system time (including satellite code biases and delays).
- c is the vacuum speed of light.
- T_r^s is the signal path delay due to the neutral atmosphere (primarily the troposphere).
- A_{IF} is the *noninteger* ambiguity of the IF carrier-phase combination, actually the IF combination of the φ_A and φ_B integer ambiguities and noninteger initial phase delays.
- λ_{IF} is the IF combination of the carrier-phase wavelengths λ_A and λ_B of signals A and B (e.g., 10.7 cm for GPS L1 and L2).
- $e_{\text{IF}}^s, \epsilon_{\text{IF}}^s$ are the relevant measurement noise components, including multipath of the IF pseudorange and carrier-phase combinations.

Since the global GNSS orbit/clock parameters are held fixed, the satellite coordinates (x^s, y^s, z^s) and the satellite clocks dt^s in (25.1) are considered known.

Furthermore, the unknown wet part of the tropospheric delay is usually expressed as a product of the wet zenith tropospheric delay ZTD_w and a mapping function that relates the slant wet delay to the zenith delay. As a result, the unknown parameters of a typical PPP model are: receiver position coordinates (x_r, y_r, z_r) , receiver clock (dt_r) , zenith troposphere delay (ZTD_w) and (noninteger) IF carrier-phase ambiguities (A_{IF}) .

After fixing the known satellite clocks and positions, the above observation equations contain observations and unknowns pertaining to a single station only. Note that satellite clock and orbit weighting does not require the satellite clock and position parameterizations, since they can be effectively accounted for by satellite-specific pseudorange/phase observation weighting. When fixing orbits/clocks, it makes little or no sense to solve (25.1) in a network solution, as it would still result in uncorrelated station solutions that are exactly identical to independent, single station, PPP solutions. Also note that, unlike relative or network solutions utilizing DD phase observations, it is not possible to fix individual integer ambiguities for the two signals A and B in single point positioning solutions without additional parameterization of measurement biases (Sect. 25.3.4).

It is worth noting that PPP provides position, ZTD and receiver-clock estimates that are consistent with the global reference system implied by the fixed global GNSS orbit/clock solutions. The DD approach, on the other hand, does not offer any clock solutions, and the ZTD solutions may be biased by a constant (datum) offset, in particular for regional or local network, or single baseline (< 500 km) solutions. This ZTD bias, in turn, may cause a small-scale error in relative height solutions. Thus, such regional or local ZTD solutions, based on DD analyses, require external tropospheric ZTD calibration (at least at one station of the network), for example by means of the IGS tropospheric combined ZTD products (Chaps. 38 and 33).

Traditionally, GPS L1 and L2 observation pairs have been used in PPP applications in view of the availability of highest precision orbit and clock products compatible with these signals. However, some GNSS, like the emerging Galileo or the modernized GPS systems, may also provide E5 or L5 carrier-phase observations instead of, or in addition to, those on the L2 frequency. The above dual-frequency PPP discussion is generically valid for any pair of (sufficiently spaced) signal frequencies f_A and f_B . The possible use and impact of three frequency observations, which are also becoming available for new or modernized GNSSs, are briefly discussed below in Sects. 25.2.1 and 25.3.4.

25.1.2 Adjustment and Quality Control

The design matrix needed for the adjustment follows from a linearization of the observation equations around the approximate parameter values (Chap. 21). It consists of the partial derivatives of (25.1) with respect to the four types of PPP parameters: station position, receiver clock, zenith troposphere delay, and (noninteger) IF carrier-phase ambiguities.

Batch versus Sequential

The adjustment can be done in a single step, the so-called batch adjustment (with iterations), or alternatively within a sequential adjustment or filter (with or without iterations) that can be adapted to varying user dynamics (Chap. 22). The disadvantage of a batch adjustment is that it may become too large even for modern and powerful computers, in particular for a very large number of undifferenced observations. However, no back substitutions or back smoothing is necessary in this case, which makes batch adjustment attractive in particular for DD approaches. Filter implementations for GNSS positioning are equivalent to sequential adjustments with steps coinciding with observation epochs. They are usually much more efficient and of smaller size than the batch adjustment implementations, at least as far as the position solutions with undifferenced observations are concerned. This is so, since parameters that appear only at a particular observation epoch, such as station clock and even ZTD parameters, can be preeliminated. However, filter (sequential adjustment) implementations require backward smoothing (back substitutions) for the parameters that are not retained from epoch to epoch (e.g., the station clock and ZTD parameters).

Furthermore, filter or sequential approaches can also model variations in the states of the parameters between observation epochs with appropriate stochastic processes that also update parameter variances from epoch to epoch. For example, the PPP observation model involves four types of parameters: station position (x_r, y_r, z_r) , receiver clock (dt_r) , troposphere zenith path delay (ZTD_w) and noninteger carrier-phase ambiguities (A_{IF}) . The station position may be constant or change over time depending on the user dynamics. These dynamics could vary from tens of meters per second in the case of a land vehicle to a few kilometers per second for a low Earth orbiter (LEO). The receiver clock may drift and will have noise characteristics according to the quality of its oscillator, for example about 0.1 ns/s (equivalent to several cm/s) in the case of an internal quartz clock with frequency stability of about 10^{-10} . Comparatively, for a stationary receiver, the tropospheric ZTD will vary in time by a relatively small

amount, in the order of a few cm/h. Finally, the carrier-phase ambiguities will remain constant as long as the satellite is not being reoriented (e.g., during an eclipsing period, see the phase wind-up correction, Chap. 19 and Sect. 25.2.2) and the carrier phases are free of cycle slips, a condition that requires close monitoring. Note that only for DD data, i. e., two satellites observed from two stations, all clocks including the receiver-clock corrections are practically eliminated by the double differencing operation.

The system or process noise can be adjusted according to user dynamics, receiver-clock characteristics and atmospheric activity. In all instances the ambiguity process noise is set to zero, since the carrier-phase ambiguities remain constant over time. In static mode, the user position is also constant and consequently the coordinate process noise is also zero. In kinematic mode, it can be increased as a function of user dynamics, though usually the coordinate process noise values are set to a very large value to accommodate all possible user dynamics (including LEO satellites), effectively forcing independent position solutions for every epoch. The receiver-clock process noise can vary as a function of its frequency stability but is usually set to white noise with a large process noise variance to accommodate the unpredictable occurrence of clock resets. A random walk process noise of about $2\text{--}5\text{ mm}/\sqrt{\text{h}}$ is usually assigned and used to drive the process noise variance of the ZTD. Note that for the most precise PPP applications, ZTD modeling typically also includes two additional stochastic (e.g., random walk) unknown parameters pertaining to the north-south and east-west ZTD gradients (Sect. 25.2.1).

Data Screening and Editing

When undifferenced code and phase observations are used, such as is the case for PPP, data testing and editing is quite essential (Chap. 24). For undifferenced, single-station observations this is a major challenge, in particular during periods of high ionospheric activity and/or station in the ionospherically disturbed subauroral or

equatorial regions. This is because the difference between the phase observations on the two frequencies (e.g., GPS L1 and L2) along with widelane pseudorange/phase combinations (Chap. 20) are commonly used to check and edit cycle slips and outliers. Under quiet ionospheric conditions it is possible to detect and correct cycle slips even for data breaks exceeding 1 min, in particular when changes in ionospheric delays are taken into account [25.4]. When it is not possible to correct cycle slips a new initial ambiguity unknown has to be introduced. However, in the extreme cases of a highly active and scintillating ionosphere, this cycle slip editing approach would need data sampling higher than 1 Hz in order to safely edit or correct cycle slips or outliers. Due to memory constraints, data cannot always be sampled or processed at a rate of 1 Hz or higher. Within a geodetic receiver, however, it should be possible (at least in principle), to do efficient and reliable data cleaning and editing based on fitting the individual carriers phase measurements (e.g., φ_{L1} and φ_{L2}) or their difference ($\varphi_{L1} - \varphi_{L2}$), since data samplings much higher than 1 Hz are internally available. Most IGS stations have data sampling of only 30 s, which is why efficient statistical editing and error detection tests are critical, in particular for undifferenced, single station observation analyses.

On the other hand, the double-difference carrier-phase observations on the individual frequencies or even the DD ionosphere-free measurement combinations are much easier to edit or correct for cycle slips and outliers, consequently making statistical error detection and corrections less critical or even unnecessary. An attractive alternative for undifferenced observation network analyses is cycle slip detection and editing based on DD observations, which could also facilitate the resolution of the initial DD phase ambiguities. Resolved phase ambiguities are then reintroduced into the undifferenced analysis as the condition equations of the new undifferenced observations, formed from the reconstructed, unambiguous and edited DD observations, previously obtained.

25.2 Precise Positioning Correction Models

GNSS software must apply corrections to pseudorange observations in order to eliminate effects such as special and general relativity, Sagnac delay, satellite clock offsets, atmospheric delays, and so on (e.g., [25.7]; Chap. 19). Since these effects are quite large, exceeding several meters, they must be considered even for pseudorange positioning at the meter precision level. When attempting to combine satellite

positions and clocks precise to a few cm with IF carrier-phase observations (with mm precision), or for the most precise differential phase processing mode, it is important to account for additional effects that are not normally considered for pseudorange positioning. An overview of the various model components and corrections in PPP applications is provided in Table 25.1.

Table 25.1 PPP a priori correction models. Magnitude and uncertainty values should be considered as approximate and may differ from case to case (after [25.27])

Model component		Magnitude	Uncertainty	Notes
Satellite	Center-of-mass position		2.5 cm (GPS)	Interpolated from precise orbit product in standard product 3 (format) (SP3) format with typical sampling of 15 min
	Antenna phase center offset	0.5–3 m	10 cm	Antenna offset vector in spacecraft system (IGS antenna exchange (ANTEX)) and GNSS specific attitude models [25.5, 6]
	Phase center variations	5–15 mm (GPS)	0.2–1 mm	IGS ANTEX model [25.5]
	Clock offset	< 1 ms	75 ps, 2 cm (GPS)	Interpolated from precise clock product with typical sampling of 30 s to 5 min
	Relativistic clock effects	10–20 m	–	Eccentricity-dependent effect [25.7, 8]
		2 cm	–	J_2 -dependent contribution [25.8]; consistently neglected in current precise GNSS clock products and PPP models
Atmosphere	Differential code biases	up to 15 ns, 5 m	0.1–1 ns	Required biases depend on tracked signals and clock product [25.9, 10]
	Fractional phase biases	up to 0.5 cy	0.01 cy	For undifferenced ambiguity resolution [25.11]
	Troposphere (dry)	2.3 m	5 mm	Vertical delay [25.12], up to 10× larger for low elevations. Models: see, e.g., [25.13, Sect. 9.2], [25.14, 15]
	Troposphere (wet)	up to 0.3 m	up to 100%	Vertical delay [25.12]; estimated due to insufficient a priori models
	Ionosphere (1 st -order)	up to 30 m	– / 1 m	Vertical delay, up to 3× larger for low elevations. Corrected through ionosphere-free combination (2-freq. PPP) or global ionosphere maps ([25.16]; 1-freq. PPP)
	Ionosphere (higher-order)	0–2 cm	1–2 mm	References [25.17] and [25.13, Sect. 9.4.1]
Site displacement				Corrections for expressing measured positions in a conventional terrestrial reference frame
	Plate motion	up to 0.1 m/y	0.3 mm/y	Reference [25.18]
	Solid Earth tide	up to 0.4 m	1 mm	References [25.19] and [25.13, Sect. 7.1.1]
	Ocean loading (tidal)	1–10 cm	1–2 mm	References [25.13, Sect. 7.1.2], [25.20, 21]
	Ocean loading (nontidal)	up to 15 mm	1 mm	Nonconventional correction; [25.22]
	Pole tide	25 mm	–	Reference [25.13, Sect. 7.1.4]
	Atmospheric loading (tidal)	up to 1.5 mm	–	Reference [25.13, Sect. 7.1.3]
	Atm. loading (nontidal)	up to 20 mm	15%	Nonconventional correction; [25.23]
Receiver	Phase center offset	5–15 cm	–	IGS ANTEX model (conventional values)
	Phase center variations	up to 3 cm	1–2 mm	IGS ANTEX model; [25.24]
Others	Phase wind-up	10 cm	see notes	Wavelength dependent; correction subject to knowledge of satellite/receiver antenna orientation; [25.25, 26]

For relative positioning at the cm-precision level and baselines of less than 100 km, all the correction terms discussed below can be safely neglected. The following sections describe additional correction terms often neglected in local relative positioning, that are, however, significant for PPP and all precise global analyses (DD or undifferenced approaches).

In the following discussion of PPP models, the correction terms have been grouped under four subsections covering propagation delays (Sect. 25.2.1),

antenna effects (Sect. 25.2.2), site displacements effects (Sect. 25.2.3) and differential code biases (Sect. 25.2.4). Furthermore, compatibility considerations are addressed in Sect. 25.2.5.

A number of the corrections listed below require positions for the Moon and Sun (e.g., for tide and attitude computations). The respective information can be obtained from readily available planetary ephemerides files [25.28, 29], or more conveniently from simple analytical formulas [25.30–32], since a relative precision

of about 1/1000 is sufficient for corrections at the mm precision level.

25.2.1 Atmospheric Propagation Delays

Propagation of radio waves through the Earth's atmosphere introduces significant delays, which must be taken into account even for GNSS positioning at the meter precision level. For a comprehensive description of GNSS signal propagation see Chap. 6. Below are summarized the propagation delay models required for the highest precision PPP and GNSS global solutions as outlined in the current IERS2010 conventions [25.13].

Higher-Order Ionospheric Delay Corrections

The IF linear combination of dual-frequency observations used in (25.1) can be subjected to cm-level systematic errors caused by the neglected higher-order ionospheric delays. The higher-order ionospheric delays are negligible with respect to pseudorange noise of about 0.1–1.0 m but need to be considered for phase observations [25.33].

Following [25.13], the higher-order ionospheric delay errors of IF carrier-phase observations can be described as

$$d\phi_{r,IF}^s = -\frac{s_2}{f_A f_B (f_A + f_B)} - \frac{s_3}{f_A^2 f_B^2}, \quad (25.2)$$

where f_A and f_B denote the two signal frequencies (Hz) used in the IF combination.

The third-order term s_3 is negligible (at the sub-mm level) for GNSS frequencies. However, for a very high intensity ionosphere (such as during peaks of solar activity cycles) an s_3 ray-bending contribution, Δs_3 , may become significant. For a given elevation E and slant total electron content (STEC), it can be approximated as

$$\Delta s_3 = b_1 \left(\frac{1}{\sqrt{1 - b_2 \cos^2(E)}} - 1 \right) \text{STEC}^2 \quad (25.3)$$

with $b_1 = 2.495 \cdot 10^8 \text{ mm MHz}^4/\text{TECU}^2$ and $b_2 = 0.8592$ [25.13, 34]. The slant total electron content can reach up to ≈ 300 TECU for a highly active ionosphere, where 1 TECU = 10^{16} electrons/m². Thus, $\Delta s_3/f^4$ can reach up to 10 mm and should be considered here, along with the second-order term s_2 , at least for the most precise GNSS solutions.

The s_2 coefficient of the second-order term can be approximated by

$$s_2 = 1.1284 \cdot 10^{12} B_p \cos(\theta) \text{STEC}, \quad (25.4)$$

where $B_p \cos(\theta)$ is the projection of the Earth's magnetic field intensity onto the satellite-station (i.e.,

satellite signal propagation) direction [25.13]. Equation (25.4) yields the value of s_2 in units of mHz³ for STEC measured in electrons/m² and B_p expressed in Tesla. The magnetic field strength required for the second-order correction can readily be obtained from models such as the international geomagnetic reference field (IGRF). Both the magnetic field B_p and the satellite station direction are taken at the piercing point on the adopted ionospheric shell (typically at a height of 450 km).

The STEC in (25.3) and (25.4) can be obtained from global ionosphere maps (GIMs) providing the vertical total electron content (VTEC) and a thin-shell mapping function (Chaps. 6 and 19). Such maps are, for example, generated by the IGS on a daily basis and distributed in the standardized ionosphere exchange format (IONEX); (Annex A). Alternatively, STEC can be evaluated from dual-frequency pseudorange-leveled carrier-phase observations after proper consideration of satellite- and receiver-specific differential code biases for the employed signals.

From (25.4) one can see that the second-order correction is highly geographically correlated, since it is a projection on the direction of the Earth's magnetic field, which is nearly the same within a wide area around the station. Furthermore, the direction of the Earth's magnetic field (mainly pointing north-south) is changing very slowly in time (nearly constant even over a decade). Therefore due to periodical changes of satellite geometry the second-order ionospheric refraction will cause small periodical errors, mainly in latitude. However, as seen from (25.4), the second-order ionospheric correction is also proportional to STEC, so it changes during the day (small at night, larger during the day). Finally, it can also be an order or even two orders of magnitude smaller (thus insignificant) during periods of low ionospheric activity than during periods of very active ionosphere.

In principle, the availability of three signal frequencies (such as GPS L1, L2, and L5) opens the possibility to eliminate the second-order ionospheric delay by an appropriate combination of the triple frequency observations [25.35]. However, in that case, because of additional biases connected with the third frequency observations, the compatibility with the standard dual-frequency solutions as well as a significant amplification of the observation noise [25.13] also need to be considered.

Tropospheric Delay Modeling

The tropospheric delay in (25.1) is commonly expressed as the product $T_r^s = M \text{ZTD}$ of an elevation-dependent mapping function M and the zenith tro-

posphere delay ZTD. For all GNSS frequencies, the tropospheric delay T_r^s of (25.1) does not depend on frequency and the ZTD amounts to about 2.3 m at sea level. The ZTD can conveniently be divided into hydrostatic (dry) and wet components. The hydrostatic delay is caused mainly by the refractivity of the dry gases in the troposphere. The water vapor refractivity is responsible for most of the wet delay. Typically the hydrostatic delay component accounts for about 90% of the total delay (Chap. 6).

The hydrostatic delay (ZTD_h) can be accurately computed a priori from surface pressure p , station latitude φ and height h , using the formula of *Saastamoinen* [25.36] as given by [25.37]

$$ZTD_h = \frac{0.0022768 \text{ m/hPa} p}{1 - 0.00266 \cos(2\varphi) - 2.8 \cdot 10^{-7} \text{ m}^{-1} h} \quad (25.5)$$

For the smaller wet zenith delay (ZTD_w), there is no reliable model to obtain an a priori value. Because measuring the wet delay using water vapor radiometers is expensive and impractical for GNSS, it is normally estimated from the data. Standard GNSS navigation, utilizing pseudorange measurements or relative positioning over short baselines of a few tens of km, require only a simple mapping function M and a single a priori ZTD. In such cases, ZTD estimation is usually unnecessary or impossible. On the other hand, PPP and precise global solutions (Chap. 34) require that the ZTD mapping function M also be separated into a hydrostatic (dry) part (M_h) and a wet part (M_w). For the most precise GNSS applications, the ZTD north and east gradients (G_N , G_E) are also used, along with a gradient mapping function M_g . More specifically, the tropospheric delays of (25.1) are parameterized as

$$T_r^s = M_h(E) ZTD_h + M_w(E) ZTD_w + M_g(E) [G_N \cos(A) + G_E \sin(A)], \quad (25.6)$$

where A is the azimuth of the satellite direction and the gradient mapping function

$$M_g(E) = \frac{1}{(\sin(E) \tan(E) + 0.0032)} \quad (25.7)$$

as suggested by [25.38] is typically used. The horizontal gradients (G_N , G_E) are needed to account for north-south atmospheric bulge and weather systems, since both can reach up to 1 mm [25.39].

Practically all the modern mapping functions use continued fractions

$$M(E, a, b, c) = \frac{1 + \frac{a}{1 + \frac{b}{1+c}}}{\sin E + \frac{a}{\sin E + \frac{b}{\sin E + c}}} \quad (25.8)$$

in terms of $\sin(E)$ as introduced by [25.40], where the coefficients a , b and c are small (< 1) constants. Different sets of coefficients (a_h , b_h , c_h) and (a_w , b_w , c_w) are required for the hydrostatic M_h and wet M_w mapping functions, respectively. Only the variation of the most significant coefficients a_h and a_w needs to be considered. The remaining and smaller coefficients (b and c) can use functional, mainly seasonal, representations.

For a self-contained PPP application with no external information input, the coefficients a_h and a_w can be obtained from global spherical harmonics expansions of mean geographical and seasonal variations, which is the case of the global mapping function (GMF) [25.14]. The more recent mapping function of the GPT2 (global pressure and temperature) model [25.41] uses global grids of mean values and mean seasonal or semiseasonal variations. The most precise PPP and GNSS applications use the Vienna mapping function 1 (VMF1) [25.42], which requires actual temporal and geographical variations of a_h and a_w , either site-specific or geographical grid files (with $2^\circ \times 2.5^\circ$ resolution). The VMF1 site-specific or grid files contain four sets of a_h and a_w coefficients per day (i.e., every 6 h), fitted to ray-tracing through the numerical weather model (NWM) of the European Centre for Medium-Range Weather Forecasts (ECMWF). The VMF1 site-specific or grid files, and alternatively those generated by the University of New Brunswick (UNB) and based on the US and Canadian NWMs [25.43], are readily available at [25.44] and [25.45], respectively.

Even though errors of the a priori ZTD_h (25.5) can be largely compensated by the ZTD_w estimation, for the most precise PPP and GNSS applications, ZTD_h needs to be known fairly accurately in order to properly separate the dry and wet ZTD mapping (25.6). According to [25.42], for a 5° elevation cutoff angle, the hydrostatic/wet mapping separation causes height errors of about one tenth of the ZTD_h error. This means that to reduce height errors below the mm level, the a priori ZTD_h has to be known at the cm-precision level, which in turn means that ZTD_h has to be based on measured pressure p , or more conveniently on p obtained from a NWM, for example the ones in the VMF1 grid files. The NWM grid files also contain ZTD_w , however, its uncertainty is at the 2 cm level, which is not sufficient for most PPP applications, and thus ZTD_w estimations are still required. Nevertheless, the NWM-based a priori ZTD_w can be used to significantly constrain ZTD_w .

estimates, which may shorten the initial PPP solution convergence.

The VMF1 and UNB grid files require spatial and temporal interpolations of a_h , a_w and ZTD_h , ZTD_w for a specific station location and epoch [25.46]. Less precise, self-contained PPP solutions can use a constant ZTD_h , or one evaluated from (25.5) for a specific station location and epoch using the global pressure and temperature (GPT) model pressure [25.15]. Alternatively it can be obtained directly from the more recent GPT2 routine. Both GPT and GPT2 are based on global averages of NWM values and their seasonal variations. [25.47] investigated GMF and GPT performance and [25.43] compared GPT2-based PPP solutions with those using the VMF1 and UNB grid mapping functions and ZTD_h . Using a constant ZTD_h instead of a GPT- or GPT2-derived one may result in significant height errors due to hydrostatic/wet mapping separation errors. This is true in particular at high latitudes with large atmospheric pressure variations, where height errors can exceed 10 mm. It is interesting to note that a constant or GPT-derived ZTD_h and GMF, and to a smaller extent also the GPT2-derived values, tend to compensate the atmospheric loading effects on heights [25.46]. This explains why prior to atmospheric loading corrections, PPP solutions utilizing constant or GPT/GPT2 a priori ZTD_h , and/or GMF/GPT2 mapping functions may show slightly better height repeatability than the more accurate gridded VMF1 PPP solutions.

25.2.2 Antenna Effects

Phase Center Offsets and Variations

The ephemerides broadcast by today's GNSS satellites provide the position of the satellite antenna for direct use within the position computation. Here, no knowledge of the spacecraft orientation and antenna accommodation is required, but the achievable accuracy is limited in accord with the needs of pseudorange-based navigation. High-accuracy orbit products for PPP applications, in contrast, are referred to the spacecraft center of mass (CoM), which is the primary reference point for the orbit modeling. However, since GNSS measurements are effectively made between the phase centers of the transmitting and receiving antennas, it is necessary to account for the CoM offset of the satellite antenna and the orientation of the offset vector in space.

Representative values of the satellite antenna phase center offsets are summarized in Table 25.2 for the various constellations. The phase centers of all the GNSS satellites are offset by about one to a few meters in the body z -coordinate direction (towards the Earth) and some are also offset in the body x -coordinate direction,

which is nominally in the plane containing the Sun, satellite and Earth.

In addition to the phase center offset in the spacecraft body frame, the orientation (attitude) of the spacecraft body relative to the terrestrial reference frame must be known to obtain the phase center position for given CoM coordinates of the GNSS satellite. Nominal attitude laws for the individual constellations and satellite types are discussed in [25.6] and Chap. 3. They allow computation of the satellite orientation for given orbital position and Sun direction and offer a good approximation of the true attitude except for short periods of noon and midnight turns during the eclipse season.

Since the assumption of a common phase center for all signals and line-of-sight directions is only approximately true for real antennas, complementary phase center variations (PCVs) need to be considered for high-precision carrier-phase modeling (Chaps. 17 and 19). Since November 5, 2006 (GPS Week 1400) the IGS has adopted calibration tables of absolute antenna PCV for both satellite and receiver antennas [25.5], which are readily available from the IGS [25.48] and updated as needed. The absolute PCV files (e.g., *igs08.atx* for consistent use with the IGS08/ITRF08 reference frame) contain PCV calibrations for all GNSS satellites and for practically all the receiver antenna models used by IGS. The receiver antenna PCV calibrations are usually based on antenna robot calibrations [25.24, 49] and include the measured phase center offsets (PCOs) together with elevation and azimuth dependent PCVs. The satellite portions of the absolute PCV file are based on solutions of several IGS analysis centers (ACs), which are consistent with the receiver antennas absolute PCVs and the IGS realization of the international reference frame.

It is advisable to use the absolute PCV convention in PPP solutions, for consistency with the orbits/clock computation process, but only when a receiver-absolute antenna PCV is available. If only a relative or no PCV calibration is available for the receiver antenna, then the nominal satellite antenna offsets and no satellite PCV should be used. PPP using absolute antenna PCV for satellite antennas with no or a relative receiver PCV may result in large (decimeter) solution errors and inconsistencies. Similarly, when orbits/clocks referred to satellite antenna phase center are generated from a network of ground stations (e.g., a commercial one) employing the same antenna types with no PCV, then PPP users with compatible antenna should not use any PCV either. However, when a user employs a different antenna than the one used to generate the orbits/clocks, the difference between the PCVs of the two antennas need to be accounted for.

Table 25.2 Antenna phase center offset from the center of mass for different types of GNSS satellites (after [25.6]). The offsets refer to IGS-specific spacecraft body axes and serve for illustration only. Satellite- and frequency-specific values used in the generation of IGS precise orbit and clock products are provided as part of the IGS ANTEX model (after [25.5])

Constellation	Type	x (m)	y (m)	z (m)
GPS	II/IIA	+0.28	0.00	+2.56
	IIR-A	0.00	0.00	+1.31
	IIR-B/M	0.00	0.00	+0.85
	IIF	+0.39	0.00	+1.60
GLONASS	M	-0.55	0.00	+2.30
	K1	0.00	0.00	+1.76
Galileo	In-orbit validation (IOV)	-0.20	0.00	+0.60
	Full operational capability (FOC)	+0.15	0.00	+1.00
BeiDou-2		+0.60	0.00	+1.10
Quasi-Zenith Satellite System (QZSS)	QZS-1	0.00	0.00	+3.20
Indian Regional Navigation Satellite System (IRNSS/NavIC)		+0.01	0.00	+1.28

Some modern receivers allow input of a receiver antenna PCV and output PCV corrected data, in such a case only the satellite antenna PCV should be considered when orbits/clocks refer to satellite centers of mass. When using the receiver independent exchange format (RINEX [25.50, Annex A]) for GNSS observations, data from receivers applying PCV corrections will report *NULLANTENNA* in the file header.

Phase Wind-Up

GNSS satellites employ right-hand circularly polarized (RHCP) electromagnetic waves for signal transmission, which means that the electric and magnetic field vectors perform a right-hand rotation about the propagation direction (Chap. 4). Other than linear polarization, the use of RHCP signals avoids restrictions on the relative orientation of the receive and transmit antenna and helps to mitigate multipath effects from reflected signals [25.51]. As a side effect, the measured carrier phase does not only change with the distance of the transmitter and receiver but also with the orientation of either of the two antennas relative to the line of sight. This is known as *phase wind-up* [25.25] and will, for example, result in a phase change by one cycle for a full rotation of the receive or transmit antenna about the boresight direction. It should be noted that only the carrier phase measurements are sensitive to wind-up effects, whereas the pseudorange observations remain unaffected (Chap. 19).

Phase wind-up effects have commonly been neglected in differential positioning applications since the effects are highly correlated for stationary receivers with a separation of less than a few hundred kilometers. For mobile receivers the phase wind-up caused by a rotation of the receiver antenna about a fixed axis is identical for all received satellites. Thus, it can partly be ab-

sorbed in the clock solution, but will give rise to a code-carrier inconsistency when processing both pseudorange and phase observations. Consideration of the user antenna orientation and the resulting phase wind-up is therefore essential for precise positioning on mobile platforms with continued attitude changes [25.26]. In particular, phase wind-up effects must be properly modeled for a toggling antenna [25.52], where the rotation vector varies over time.

Even for a presumably stationary position and alignment of the receiver antenna, phase wind-up effects arise from the slowly changing relative orientation of the satellite antenna, line of sight, and receiver antenna. Following [25.25] the resulting carrier phase change may differ by up to 4 cm for two stations separated by 4000 km.

GNSS satellites need to continuously change their orientation about the Earth-pointing antenna axis to orient their solar panels towards the Sun. Irrespective of the user antenna dynamics, these satellite attitude changes will result in a measurable phase wind-up effect. They are most pronounced during noon and midnight turns in the eclipse season, where the satellites may rotate by up to 180° (corresponding to a phase wind-up effect of half a wavelength) in 15–30 min. If ignored, these are fully absorbed into the estimated satellite clocks and thus are completely eliminated by double differencing. They become important, however, for undifferenced PPP applications and need to be consistently handled in the generation of orbit/clock products and the user positioning software. Within the IGS, phase wind-up effects are considered by all analysis centers and their respective products. Neglecting them and fixing IGS orbits/clocks in a PPP process may result in position and clock errors at the dm level.

Details of the phase wind-up modeling and the applicable satellite attitude models are provided in Chap. 19. Aside from nominal attitude laws, dedicated models have been developed for describing the noon or midnight turns of various types of satellites during the eclipse season [25.53–55]. Unless these models can be consistently applied by the user, the respective satellites and time intervals should be discarded in the PPP processing.

As an alternative to rigorous phase wind-up modeling, the use of a *decoupled clock model* has been suggested in [25.56]. Here unmodeled wind-up effects that otherwise result in a code-carrier inconsistency are absorbed in distinct clock offset parameters for the pseudorange and carrier-phase observation model. This approach can be applied if external attitude information for the receiver antenna is not available and the antenna is primarily rotating about a constant axis.

25.2.3 Site Displacement Effects

By its very nature, precise point positioning delivers coordinates in a global terrestrial reference frame such as the international terrestrial reference frame (ITRF) or the IGS-specific IGSy frame. The realization of such a frame is complicated by the fact that the Earth and its crust are not perfectly solid. The various forces acting on the Earth (e.g., lunar and solar gravity, but also loading due to ice, oceans and even the atmosphere) result in periodic deformations and thus periodic motion of individual stations. These are mostly highly correlated over large areas and can therefore be neglected in relative positioning over up to a few hundred km. However, the periodic motions have been removed through relevant models in the realization of the ITRF and its reference station coordinates. In accord with current IERS conventions [25.13], the same models of the periodic site displacements must be accounted in all PPP applications to obtain ITRF-compatible site positions.

Dominant effects such as solid Earth and pole tides or ocean loading cause site displacements at the few cm to dm level and are discussed below in further detail. Effects with a magnitude of less than 1 cm, such as surface loading from atmospheric pressure, ground water and/or snow buildup, are neglected and not considered in the following. These small effects can be applied a posteriori or even monitored with PPP solutions (e.g., local ground water/snow buildup variations). For these reasons, no IGS solutions currently include the above-mentioned environmental loading effects. Furthermore, diurnal and semidiurnal atmospheric tides S_1 and S_2 , included in the IERS2010 conventions and applied by some IGS analysis centers have also been neglected here. The vertical amplitudes of S_1 and S_2 can reach

up to about 2 mm, mainly in the equatorial regions, and they will largely average out over the standard 24 h solution periods used by IGS. The horizontal S_1/S_2 effects are about one order of magnitude smaller, so for all kinematic and most static PPP solutions, the horizontal and even vertical atmospheric tides can be neglected.

Solid Earth Tides

Similar to ocean tides, the gravitational attraction of the Sun and Moon causes a subtle deformation of the (presumably solid) Earth and its crust. It results in horizontal and vertical displacements that can be modeled by a spherical harmonics expansion and associated physical parameters (known as Love and Shida numbers), which describe the susceptibility of the Earth's body to the tide-generating potential. At an accuracy level of about 5 mm, it is sufficient to only consider the dominant, second-degree tides of the Sun and Moon along with a supplementary height correction term [25.57]. Within this approximation, the site displacement of a station at position \mathbf{r} can conveniently be described by the geocentric unit vectors $\mathbf{e}_\odot = \mathbf{r}_\odot/r_\odot$, $\mathbf{e}_\text{c} = \mathbf{r}_\text{c}/r_\text{c}$, and $\mathbf{e} = \mathbf{r}/r$ in Sun, Moon, and station direction

$$\begin{aligned} \Delta \mathbf{r} = \sum_{j=\odot, \text{c}} \frac{GM_j}{GM_\oplus} \frac{r^4}{R_j^3} \left\{ [3l_2 (\mathbf{e}_j \cdot \mathbf{e})] \mathbf{e}_j \right. \\ \left. + \left[3 \left(\frac{h_2}{2} - l_2 \right) (\mathbf{e}_j \cdot \mathbf{e})^2 - \frac{h_2}{2} \right] \mathbf{e} \right\} \\ + [-0.025 \text{ m} \sin \varphi \cos \varphi \sin (\theta_g + \lambda)] \mathbf{e}. \end{aligned} \quad (25.9)$$

Here, GM_\oplus , GM_\odot , GM_c are the gravitational coefficients of the Earth, Sun, and Moon, while $l_2 = 0.6090$ and $h_2 = 0.850$ are the nominal second-degree Love and Shida numbers. The height correction term in (25.9) is described in terms of the station latitude φ and longitude λ as well as the Greenwich mean sidereal time θ_g .

For an accuracy of 1 mm or better further harmonics and the dependence of Love and Shida numbers on the station location and the frequency of each tidal constituent need to be considered [25.13, 58]. To facilitate a consistent application of the respective corrections, suitable computer implementations are made available along with the IERS conventions [25.13].

Overall, the solid Earth tides induce vertical station displacements of about 0.3 m and horizontal displacements of about 5 cm. Aside from periodic contributions with a dominating half-daily and daily periodicity, the tidal correction (25.9) also comprises a permanent displacement at the 1 dm level. Even though the periodic terms are largely averaged out in the processing of daily

arcs for static sites, the same does not apply for the permanent tidal displacement. Irrespective of the data arc and type of site, consideration of the full solid Earth tide correction is therefore essential in all PPP applications to comply with the *tide-free* ITRF realization.

Rotational Deformation due to Polar Motion (Pole Tide)

Aside from luni-solar tidal forces, small periodic changes in the deformation of the Earth are also caused by polar motion, that is, by changes in the location of the Earth's rotation axis relative to its crust. Following [25.13], the associated site displacements in east, north and up direction are given by

$$\begin{aligned}\Delta r_E &= +9 \text{ mm} \cos \theta [m_1 \sin \lambda - m_2 \cos \lambda], \\ \Delta r_N &= +9 \text{ mm} \cos 2\theta [m_1 \cos \lambda + m_2 \sin \lambda], \\ \Delta r_U &= -33 \text{ mm} \sin 2\theta [m_1 \cos \lambda + m_2 \sin \lambda],\end{aligned}\quad (25.10)$$

for a station at longitude λ and colatitude $\theta = \pi/2 - \varphi$. Here $m_1 = (x_p - \bar{x}_p)$ and $m_2 = -(y_p - \bar{y}_p)$ (expressed in [']) are the coordinates of the Earth's rotation pole in the terrestrial reference frame, which are obtained as the difference of the polar motion variables (x_p, y_p) and the IERS model [25.13, Table 7.7] of the mean pole ($\bar{x}_p, -\bar{y}_p$).

Polar motion is not predictable, but exhibits dominating variations with periodicities of about 430 d (Chandler period) and 365 d (annual period). At amplitudes of up to 0.8'', the site displacements due to the pole tide may amount to roughly 25 mm in the vertical direction and about one quarter of this value in horizontal direction.

Polar motion centrifugal effects on the oceans cause an analogous ocean pole tide loading, also considered in the IERS2010 conventions. It also has seasonal and Chandler period variation, but it is rather small, nearly an order magnitude smaller than the above polar tides, so it can be safely neglected in most PPP applications.

Ocean Loading

Ocean tides result in a varying load of sea water and associated deformations of the Earth's crust. The induced site displacement is most pronounced in the vertical direction and typically at the cm level. However, in coastal regions, ocean loading can result in coordinate changes of up to 10 cm [25.20]. The response of the Earth's surface to the load changes depends largely on the topography and is typically not aligned with the body tides [25.59]. As with solid Earth tides, ocean loading effects show predominant semidiurnal and diurnal periodicities but, by convention, do not exhibit a permanent part.

Given these characteristics, ocean loading may be neglected for static positioning over daily periods, stations far off (typically > 1000 km) the coast or moderate accuracy requirements. However, it clearly needs to be considered for kinematic positioning, cm-level accuracy and coastal regions. As pointed out by [25.60], unmodeled ocean loading effects may also contaminate tropospheric ZTD or station clock estimates, which are highly correlated with the vertical position.

In its most basic form, coordinate shifts Δc due to ocean loadings are described as a harmonic series

$$\Delta c = \sum_{j=1}^{11} A_{cj} \cos(\chi_j(t) - \phi_{cj}) \quad (25.11)$$

for each of three coordinate axes [25.13]. The individual terms considered in these series correspond to one of 11 semidiurnal (M_2, S_2, N_2, K_2), diurnal (K_1, O_1, P_1, Q_1), and long-period (M_t, M_m and S_{sa}) tide waves. The time-dependent angles χ_j are linear combinations of fundamental astronomical arguments such as the mean longitudes of the Sun and Moon and can consistently be computed using reference software implementations provided by the IERS [25.13]. The amplitudes A_{cj} and phases ϕ_{cj} , on the other hand, are station-specific quantities computed from global ocean tide models [25.61]. For a specific site and ocean tide model, these values can be conveniently obtained from an ocean loading provider service [25.62].

The ocean loading also induces periodic tidal variations of the Earth's center of mass (CoM) relative to a crust-fixed system aligned with the mean center of the Earth. These CoM offsets may be evaluated using an expression similar to (25.11) [25.13], but are not normally required for PPP users, since GNSS orbits products as provided by the IGS are, by convention, referred to a crust-fixed frame such as the ITRF.

25.2.4 Differential Code Biases

The observation model discussed in Sect. 25.1.1 is based on the simplifying assumption that all measurements are free of any biases. While this assumption is not necessarily true, it offers a proper model in practice, provided that GNSS clock products are generated with the same type of observations as used for the precise point positioning. For GPS, published clock offsets (in both the broadcast ephemerides and the precise products) are conventionally referred to an ionosphere-free combination of P(Y)-code observations on the L1 and L2 frequencies. Similarly, GLONASS precise clock products are based on L1/L2 P-code observations. When using the same signals for PPP, no further code

biases need to be considered and the observation model (25.1) can be used as is.

The situation is different, though, when working with other types of dual-frequency signals (e.g., the civil L1 C/A or L2C codes). In this case satellite-specific differential code biases (DCB) have to be applied to account for group delay differences between the signals tracked by the receiver and those of the clock reference signal [25.63]. A common application case is dual-frequency GPS PPP using commercial receivers that do not provide distinct P(Y)-code observations on L1, but deliver only C/A-code pseudoranges. Here, a supplementary bias

$$\frac{f_{L1}^2}{f_{L1}^2 - f_{L2}^2} \text{DCB}_{\text{CIC-C1W}}^s \quad (25.12)$$

needs to be added in the observation model (25.1) for the ionosphere-free pseudorange to translate the satellite clock offset and make it compatible with the employed observations. In the above equation

$$\text{DCB}_{\text{CIC-C1W}}^s = d_{\text{CIC}}^s - d_{\text{C1W}}^s \quad (25.13)$$

denotes the differential code bias of L1 C/A and L1 P(Y) pseudorange observations (indicated here by the corresponding RINEX observation codes CIC and C1W [25.50]). It may be noted that no receiver biases need to be considered in single-constellation processing, since those can readily be absorbed in the receiver-clock bias estimate. As an exception, such biases need to be calibrated and taken into account in PPP-based time transfer as further discussed in Chap. 41.

In multi-GNSS processing, satellite-specific DCBs need to be individually considered for a constellation whenever the tracked signals are different from the clock reference signals. In addition, an intersystem bias needs to be adjusted to compensate for time system differences between constellations and receiver-specific differential code biases (Chap. 21 and [25.63]).

DCBs of GPS and GLONASS satellites are routinely determined by various IGS analysis centers as part of their ionospheric analysis [25.16] for the legacy signals on L1 and L2. DCBs for the multitude of new signals and constellations are, furthermore, determined by the IGS from observed code differences and global ionosphere maps [25.10]. Use of these biases assists a more rigorous modeling of pseudorange ob-

servations. Even though PPP performance is generally driven by the high precision of carrier-phase observations, and partly *tolerant* to pseudorange errors, the proper consideration of DCBs is known to improve the convergence time in filter-based implementations and to enable a faster and more reliable ambiguity fixing.

25.2.5 Compatibility and Conventions

Precise point positioning fixes (or tightly constrains) external data such as the GNSS orbits and clock offset values. To ensure the desired cm- or mm-level accuracy, the PPP models and algorithms must be highly consistent with those used in the generation of the auxiliary products. Since PPP is in fact equivalent to a station position solution within a global network solution (but conveniently condensed within the precise orbit/clock products), it must adhere to the same conventions used in extracting orbit and clock data from the network. Among others, this may affect the choice of reference frames, Earth orientation parameters, antenna offset and phase pattern or the application of specific model corrections.

Within the IGS, which serves as a primary source of freely available high-accuracy GNSS data and products for scientific users, orbit and clock products are generated by various analysis centers (ACs). These adhere to common standards such as the IERS conventions (currently [25.13]), reference frames (currently ITRF2008/IGS08), and antenna phase center calibration models (currently *igs08.atx*, [25.48]). Clock products for GPS and GLONASS are based on ionosphere-free combinations of P(Y)-or P-code observations on the L1 and L2 frequencies and have been corrected for the eccentricity-dependent periodic relativistic clock variation. For other constellations, initial products provided within the IGS multi-GNSS experiment (MGEX) [25.64] are based on ionosphere-free E1/E5b (Galileo) or B1/B2 (BeiDou) combinations, but no formal standard has been established yet.

An overview of past and current conventions for the use of IGS products in PPP applications is provided in [25.65]. For specific and detailed information in a standardized format on each IGS AC global solution strategy, modeling and departures from the conventions, refer to the IGS central bureau archives [25.66].

25.3 Specific Processing Aspects

The concept of precise point positioning was originally developed for use with dual-frequency GPS observations, but is highly generic and can be applied to a variety of signals and constellations. Even though the basic modeling techniques discussed before are valid for all forms of PPP, some variants deserve specific consideration. Within the present section, single-frequency PPP is first discussed (Sect. 25.3.1), which is of particular interest for use with low-cost GNSS receivers. The use of GLONASS observations brings the added complexity of channel-dependent biases and is addressed in Sect. 25.3.2, while the use of new signals and other constellations is discussed in Sect. 25.3.3. Finally, PPP ambiguity fixing concepts are presented in Sect. 25.3.4, which offer a substantial increase in accuracy as well as a notably improved convergence time in sequential processing.

25.3.1 Single-Frequency Positioning

Traditional single-frequency point positioning (PP) utilizes pseudoranges only. If carrier-phase observations are available, they are commonly used for smoothing of pseudoranges, often internally within the receiver, in order to reduce pseudorange measurement noise [25.67]. The phase-smoothed pseudoranges are then used, along with ionospheric models (e.g., the broadcast Klobuchar model or global ionosphere maps [25.16]) to account for significant ionospheric delays [25.68–70]. Single frequency PP can use either the broadcast or more precise post-mission orbit/clock solutions. The broadcast and precise orbit/clocks are typically determined from dual-frequency data, so the satellite clocks reflect the corresponding differential code biases (e.g., the $DCB_{C1W-C2W}$ of GPS P(Y)-code pseudoranges on the L1 and L2 frequencies), which change from satellite to satellite and can reach several meters. Since single-frequency GPS receivers most commonly track the civil C/A-code signal rather than the encrypted P(Y)-code, an additional $DCB_{C1C-C1W}$ must be considered as well. For real-time use, equivalent timing group delay (TGD) and intersignal correction (ISC) parameters are transmitted as part of the modernized GPS navigation message [25.63, 71]. Neglecting these biases can cause positioning errors larger than when ionospheric delays are neglected [25.72]. The traditional, single-frequency PP is typically used for m-level navigation solutions only with four unknowns (three position coordinates and one clock). In such PP solutions, except for antenna offsets and a nominal tropospheric delay, practically all the effects discussed in Sect. 25.2 can be safely neglected.

A more precise alternative to single-frequency pseudorange PP is single-frequency PPP utilizing the code-plus-carrier (CPC) or GRAPHIC (group and phase ionospheric calibration [25.73]) combination $\phi_{GPH} = (p + \varphi)/2$ of pseudorange and phase observations on the same frequency. It is ionosphere-free (to first order), since the ionospheric code and phase delays are the same, but of opposite signs. Namely, the carrier phases are advanced (shortened) and pseudoranges are delayed (lengthened) by the ionosphere (Chap. 19). Consequently, the new observable has a significantly lower observation noise (by a factor of two) than the original pseudorange and requires no external ionospheric information or corrections. However, due to the use of carrier phases, is subject to an ambiguity. This necessitates the use of pseudoranges and solving for ambiguities, much like in case of the standard dual-frequency PPP. This also results in a fairly long solution convergence (15 min or longer).

Many of the models and effects discussed above (Sect. 25.2) also need to be considered here, since the GRAPHIC-based PPP precision is at a few dm (Fig. 25.1). This applies specifically to tidal site displacement effects but also to carrier-phase wind-up when working with rotating and tumbling platforms [25.26]. As with the dual-frequency carrier-phase combination, the code-plus-carrier combination is not rigorously ionosphere-free but likewise includes some second- (and third-) order contributions. These are typically buried in the observation noise and multipath but may be taken into account in case of high ionospheric activity and when working with high-performance ranging signals.

The observation model for the single-frequency GRAPHIC combination is given by

$$\frac{1}{2} (p_r^s + \varphi_r^s) = \rho_r^s + c (dt_r - dt^s) + T_r^s + \lambda A_{GPH} + \frac{1}{2} \lambda \omega + e_{GPH}, \quad (25.14)$$

where ω denotes the phase wind-up effect and $A_{GPH} \approx -N/2$ is the (float valued) GRAPHIC ambiguity. It lumps minus one half of the carrier phase ambiguity as well as differential code biases between the employed single-frequency pseudorange observation and the (dual-frequency) code observations used for the satellite clock product. These DCBs need also be considered when combining the GRAPHIC observations with (single-frequency) pseudoranges to enable estimation of both the receiver clock offset and the GRAPHIC ambiguities. Depending on the accuracy

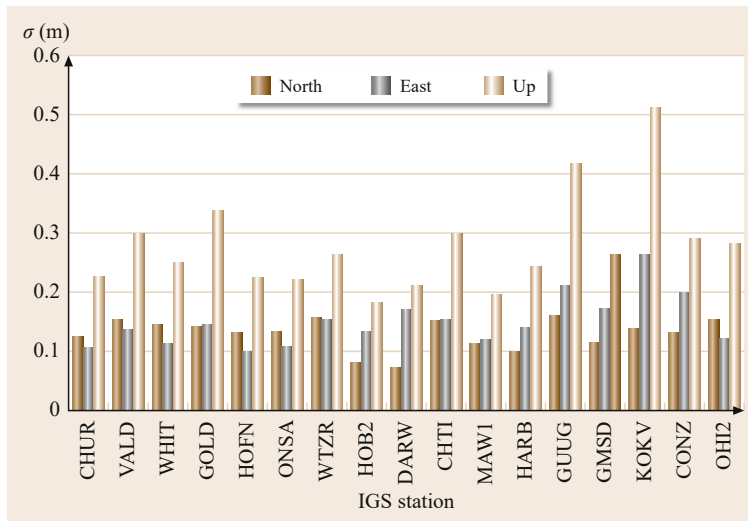


Fig. 25.1 Repeatability of single-frequency kinematic PPP solutions at 17 globally distributed IGS stations obtained with IGS final orbits/clocks over a one-year period (Jan. 1, 2012–Feb. 9, 2013). The employed GRAPHIC observations are based on GPS L1 P(Y)-code and phase measurements. Individual bars for each station indicate the precision of the north, east, and up (height) components

requirements, tropospheric delays in (25.14) may be considered through models or estimated as in the case of dual-frequency PPP [25.74].

Except for very short data arcs that do not enable proper estimation of the ambiguities, the GRAPHIC-based PPP solution usually offers better positioning results than single-frequency pseudorange processing with a priori corrections from global ionosphere maps [25.75]. As discussed in [25.76], sub-decimeter accuracy (3-D RMS) can be achieved for least-squares solutions using batches of at least 6 h duration solutions. The ionosphere-free code-plus-carrier combination appears particularly attractive for use with advanced ranging signals such as the Galileo E5 alternative BOC (AltBOC) signal. Even though the GRAPHIC processing is not fully competitive to dual-frequency PPP, a three- to four-fold performance increase has been demonstrated in [25.77] when using AltBOC in comparison with legacy GPS L1 C/A observations. Due to the very low noise and high multipath resistance of this signal, 3-D RMS positioning accuracies of 20 cm down to 3 cm can be achieved with data arcs of 1–24 h.

As an alternative to the ionosphere-free code-plus-carrier processing, the direct processing of pseudorange and carrier-phase estimation has been proposed in [25.78] along with the estimation of the vertical total electron content (VTEC) and a common mapping function. Due to different pierce points a common VTEC for all observations is not appropriate though, and horizontal ionospheric gradients have to be estimated as well in this approach.

Irrespective of the specific formulation of the single-frequency PPP algorithms, a reliable detection and handling of cycle slips is vital for achieving a high overall performance. Since GRAPHIC observations may

exhibit a noise level above the (half) carrier-phase wavelength, single-cycle slips may be hard to identify on this combination alone. Cycle-slip detection (and repair) techniques based on time-differenced carrier-phase observations and a geometry-based approach that overcome these difficulties are discussed in [25.79].

25.3.2 GLONASS PPP Considerations

Next to GPS, the Russian GLONASS was the second global navigation satellites system considered for precise point positioning [25.80, 81]. A joint use of both constellations promises notably improved convergence times and robustness, even if the accuracy of the estimated positions remains similar to that of GPS-only solutions [25.82–84].

However, the processing of GLONASS observations is complicated by the frequency division multiple access (FDMA) modulation scheme (Chap. 8), which makes use of slightly different signal frequencies on about 15 distinct channels. The individual channels are separated by 562.5 kHz and 437.5 kHz for L1 and L2, respectively (Chap. 8) and may result in interfrequency-channel biases (IFCB) for both code and phase observations. These biases affect the generation of precise orbit and clock products as well as the use of GLONASS observations for precise point positioning.

Depending on the receiver design, notable group delay variations across the different frequency channels may be encountered. Receiver-specific pseudorange IFCBs can exceed ± 10 m and usually tend to have a linear behavior with rates up to about ± 2 ns per channel index. The biases tend to be similar for the same receiver type, though antenna model and receiver model, or even a different receiver firmware version may

cause atypical behavior (Fig. 25.2). When GLONASS pseudoranges are weighted with sufficiently large uncertainty (e.g., 10 m), the pseudorange IFCBs have no significant effect (i.e., sub-mm) on PPP position and ZTD solutions but affect the receiver clock solutions. So in principle, pseudorange IFCBs need not cause significant problems in PPP solutions, unless ambiguity fixing is attempted (Sect. 25.3.4).

Aside from pseudorange IFCBs, GLONASS observations are also affected by carrier-phase IFCBs. These may differ by up to 5 cm per channel index between different receiver brands [25.85] and notably affect the ambiguity resolution in both differential and undifferenced (PPP) processing schemes. However, those small frequency-dependent phase biases are largely deterministic and can be attributed to group delays and digital delays in the signal processing that differ between receivers. They can essentially be eliminated when pseudorange and phase observations are made at the same sampling epoch [25.86].

To cope with these issues and to facilitate a consistent processing of GNSS observations from different receivers, current versions of the RINEX standard [25.50] specify a mandatory phase alignment of the GNSS measurements prior to generating the RINEX observations file. Consequently, for properly generated RINEX GLONASS data there should not be any phase IFCBs.

Provided that all the GNSS processes are using consistent and sufficiently precise observations and modeling, apart from distinct clocks solutions, each GNSS-specific PPP solution should then yield statistically equivalent position and ZTD solutions. This is confirmed, for example, in a performance assessment of [25.84] and independently illustrated in Fig. 25.3. The figure provides a comparison of GPS

and GLONASS PPP solutions for selected IGS reference stations over a 13-month period. Differences between the GPS-only solution and the GLONASS-only solutions as well as deviations from the known IGS08 positions of the stations are generally at the few-mm level and most mean offsets are statistically insignificant when the real accuracy of PPP is considered (Sect. 25.5).

It is interesting, though, to note that GNSS-specific PPP solutions are fairly independent despite the fact that the measurements are observed by the same instrumentation (receiver/antenna). This is due to different observation sets, different satellite modeling and even local environmental effects (such as multipath and subdaily station movements), which may be somewhat different due to different constellation-specific satellite geometry, signal strength and/or frequencies. For example, GLONASS-GPS daily static PPP position solution differences at coastal stations may exhibit significant fortnight periodical signals (exceeding the repeatability sigmas) when ocean loading is neglected or wrongly applied. Satellite geometry and its repeatability likely cause this, since they are different for GPS and GLONASS. In this regard, different GNSSs may facilitate an important verification of individual PPP solutions.

25.3.3 New Signals and Constellations

The ongoing modernization of the GPS and GLONASS constellations as well as the buildup of new global and regional navigation satellite systems (BeiDou, Galileo, QZSS, IRNSS/NavIC) offers new prospects for improved PPP performance but poses also a variety of challenges to their users.

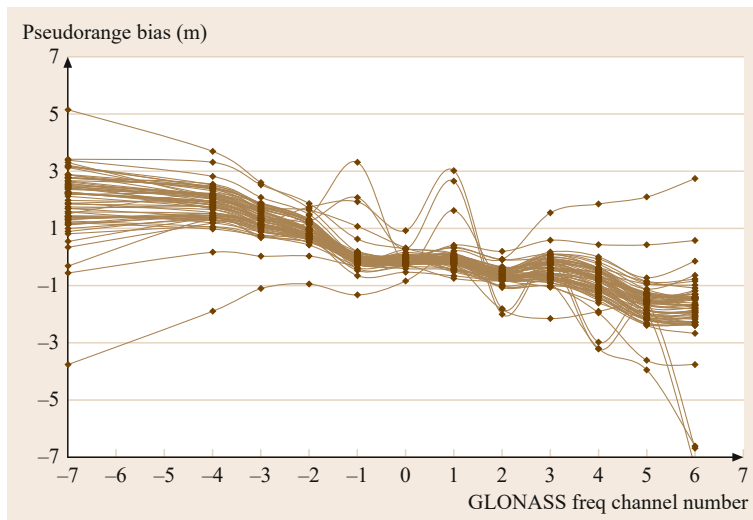


Fig. 25.2 GLONASS interchannel pseudorange biases for a group of Leica receivers as determined on March 1, 2013. Most atypical biases seen here are due to different antennae or old receiver firmware

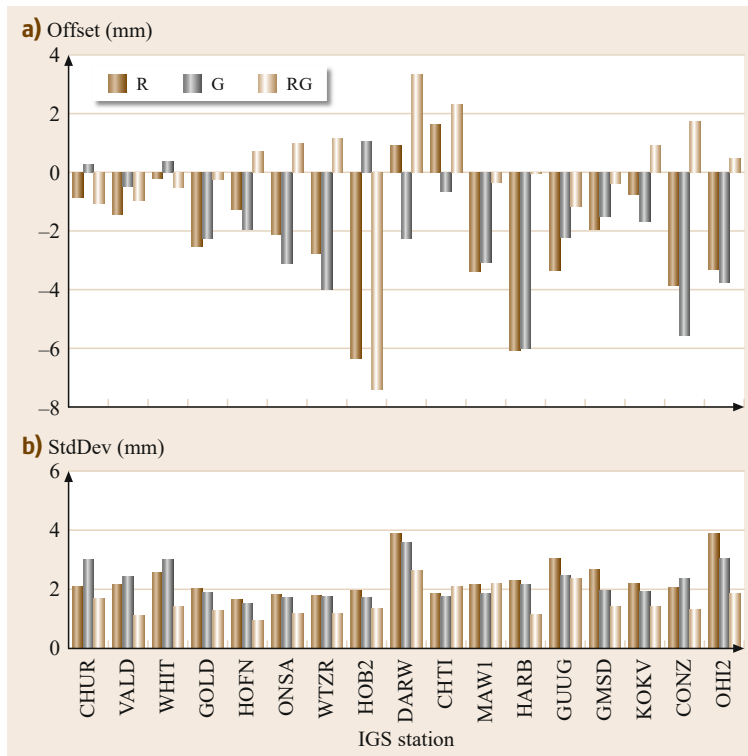


Fig. 25.3a,b Performance of static daily PPP solutions using GLONASS and GPS observations for 17 globally distributed IGS stations between 1 Jan. 2012 and 9 Feb. 2013 using European Space Agency (ESA) final GPS/GLONASS orbit/clock products. The graphs show the mean offset (a) and standard deviation (b) of the north position component for GLONASS-only processing (R) and GPS-only processing (G) relative to the IGS08 reference position as well as the difference of the two solutions (RG). Similar but slightly larger values apply for the scatter of the east and height components but no significant biases are obtained (not shown)

As discussed before, PPP depends critically on the consistency of auxiliary products (specifically the GNSS orbits and clocks) and the user processing. While relevant standards and conventions have evolved for GPS and GLONASS legacy signals over many years, they still need to be established for new signals and constellations. Along with that comes a need to thoroughly characterize the space segment (satellites, attitude laws, transmit antennas, biases) and the user segment (receivers, antennas, biases) in order to fully exploit the performance offered by the multitude of new signals in space.

While consistency of processing schemes, algorithms and even equipment can readily be ensured by commercial PPP service providers taking care of both the orbit/clock product generation and their usage, a larger effort is required for public services such as the IGS, which need to deal with a variety of different end-user equipment and possible processing tools. Such work has been initiated by the IGS within its Multi-GNSS experiment (MGEX; [25.64]) and resulted in the evolution of standards for the real-time and offline exchange of observation and navigation data (RINEX, RTCM; Annex A), conventional attitude models for all GNSS satellites [25.6], multifrequency receiver antenna calibrations [25.87], differential code biases for open signals of the various constellations [25.10] as well

as early orbit and clock products for Galileo [25.88], BeiDou [25.89], and QZSS [25.90] or several of these new constellations [25.91, 92]. Even though the precision and accuracy of multi-GNSS products and system characterizations still lags behind GPS and GLONASS, continued efforts are made to improve their performance and to make them fully competitive.

As a straightforward extension of the GPS-only or GPS-GLONASS PPP concept discussed before, ionosphere-free combinations of pseudorange and phase observations of dual-frequency signals may be processed for any individual constellation or any combination of two or more constellations. When combining signals from multiple constellations, an intersystem bias (assumed to be constant over the processing arc) needs to be estimated for all but one constellation to compensate for possible system time offsets and constellation-specific receiver biases ([25.63, 93–96] and Chap. 21). Furthermore, satellite-specific DCBs will need to be applied, if the employed signals differ from those used in the generation of the respective clock product. The choice of signals used for the individual constellations will depend on availability (all satellites in the constellation should transmit the selected signal to avoid the presence of additional biases), signal characteristics (C/N_0 , multipath resistance, etc.), and the employed clock product. Obviously, the two frequencies should be

widely spaced to minimize the noise of the ionosphere-free combination (allowing, e.g., GPS L1/L2 or L1/L5, but ruling out GPS L2/L5 as a meaningful option).

Initial results of multi-GNSS PPP processing involving BeiDou and/or Galileo next to GPS and GLONASS have, for example, been reported in [25.92, 97–99]. They confirm the benefit of an increased number of signals in space for the robustness and convergence time of PPP solutions and demonstrate an improved accuracy when applying state-of-the-art models. The combination of signals from multiple constellations is of particular interest in constrained environments, which inhibit the use of low-elevation signals. Here, the minimum number of satellites required for kinematic point positioning (four to seven depending on the number of constellations and constellation-specific intersystem biases) can be ensured for cutoff angles as high as 40° in the combined GPS, GLONASS, BeiDou and Galileo service area [25.100]. Even though other constellations than GPS and GLONASS do not yet offer a fully global availability, multi-GNSS PPP is an emerging trend that will help to further improve PPP performance but also enables a better understanding of possible systematic errors that may go undetected in single-system solutions.

Along with the integration of the new constellations into the traditional, dual-frequency PPP concept, efforts are made to allow a seamless use of signals on more than just two frequencies into the PPP processing. This is of interest, since civil (or at least publicly accessible) signals are made available on three or even more frequencies by various new or modernized navigation satellite systems (including, so far, GPS L1/L2/L5, Galileo E1/E5a/E5b/E6, QZSS L1/L2/L5/E6, and BeiDou B1/B2/B3). A possible approach consists of the joint processing of multiple ionosphere-free dual-frequency combinations (e.g., GPS L1/L2 as well as GPS L1/L5). However, special care needs to be taken in this case to account for the correlation introduced by the repeated use of the same measurements (here L1) in the combined observations [25.101].

For a unified treatment of multiple signals, an uncombined, or *raw*, processing approach is followed in [25.102–104], which uses uncombined code and phase measurements on each of the available frequencies and introduces ionospheric slant delays as additional (epoch-wise) estimation parameters. With an undifferenced formulation one has the advantages of being able to use the simplest observational variance matrix and having all the parameters remain available for a possible further model strengthening. This latter aspect allows one to take advantage for instance of the time stability of biases or next-generation satellite

clocks. Parameters that are not considered of interest can then easily be eliminated through the reduction of the normal equations, instead of performing an a priori elimination at the observational level that usually comes at the expense of a more complicated structure of the observational variance matrix. So far, experience with the *raw* PPP approach is limited due to the small number of satellites transmitting triple-frequency signals as well as time-varying biases between the L1, L2, and L5 signals of the GPS Block IIF satellites [25.105], which inhibit a proper exploitation of this method. This experience will grow with the advent of more signals and satellites, thus allowing a proper assessment of the different approaches.

25.3.4 Phase Ambiguity Fixing in PPP

Two principal benefits arise from fixing ambiguities to integers in the PPP context: improved positioning accuracy, specifically in the east component, and for filter-based PPP implementations, a reduction or possible elimination of the initial PPP solution convergence period. The latter benefit is particularly sought for the delivery of real-time PPP services, increasing their efficiency in achieving optimal solution accuracy.

DD phase ambiguity fixing has matured and is now routinely applied (Chap. 23) in either global or local positioning solutions. However, this is not directly applicable to PPP, due to the presence of pseudorange biases d and carrier-phase biases δ , which are eliminated in DD processing. This can be seen through a reparameterization of the basic observation model (25.1) explicitly exposing the different biases as follows

$$\begin{aligned}
 p_{r,A}^s &= \rho_r^s + c(dt_r - dt^s) + c(d_{r,A} - d_A^s) \\
 &\quad + T_r^s + I_r^s + e_A, \\
 p_{r,B}^s &= \rho_r^s + c(dt_r - dt^s) + c(d_{r,B} - d_B^s) \\
 &\quad + T_r^s + \mu I_r^s + e_B, \\
 \phi_{r,A}^s &= \rho_r^s + c(dt_r - dt^s) + c(\delta_{r,A} - \delta_A^s) \\
 &\quad + T_r^s - I_r^s + N_A \lambda_A + \epsilon_A, \\
 \phi_{r,B}^s &= \rho_r^s + c(dt_r - dt^s) + c(\delta_{r,B} - \delta_B^s) \\
 &\quad + T_r^s - \mu I_r^s + N_B \lambda_B + \epsilon_B.
 \end{aligned} \tag{25.15}$$

The measurement biases affecting undifferenced observations are functionally indistinguishable from the clock and ambiguity parameters. Disregarding them in PPP solutions, whether they originate from pseudorange or carrier phases, contaminates clocks and ambiguities. In recent years, much attention has been given by different research groups to resolve this issue [25.102, 106–113]. The proposed solutions require external information, in addition to the usual satellite

orbit/clock products, to break the link between biases and ambiguities and to restore the integer nature of the ambiguities. The proposed methods are largely equivalent as their differences lie primarily in the chosen parameterizations, in the way the rank deficiencies are eliminated and whether or not they make use of the ionosphere-free combined observations [25.11].

The decoupled clock model (DCM) described in [25.109, 114] combines the four observations of (25.15) into three combinations: the two ionosphere-free (IF) combined pseudoranges and carrier phases and the Melbourne–Wübbena combination (Chap. 20). The IF combined observations each have specific satellite and station clock parameters ($dt_{r,PIF}$, dt_{PIF}^s , $dt_{r,\varphi IF}$, $dt_{\varphi IF}^s$), which include the respective combined biases ($d_{r,IF}$, d_{IF}^s , $\delta_{r,IF}$, δ_{IF}^s). The Melbourne–Wübbena combination is parameterized with the usual widelane–narrowlane (WL/NL) ambiguities and station and satellite WL biases ($d_{r,WL}$; d_{WL}^s). All IF and WL biases are combinations of the original observation biases. In the DCM implementation, the integer nature of ambiguities is assured by fixing a minimum set of ambiguities to arbitrary integers (the ambiguity datum), while estimating the clock and bias parameters at discrete epochs. The N_{WL} and N_{NL} can then be resolved using common integer search schemes (Chap. 23). The additional satellite parameters required for PPP under DCM are, apart from the clock corrections, which now are pseudorange-specific and carrier-phase-specific, also satellite specific WL biases. The PPP algorithm must estimate two station clocks, one for each observation type, and a station WL bias. Finally, an ambiguity datum must be maintained within the PPP through a minimum set of ambiguities fixed to arbitrary integers.

The integer recovery clock (IRC) approach described in [25.115] uses the same base observation combinations (two IF and one WL), however the parameterization is slightly different: in addition to the satellite and station WL biases and ambiguous phase clocks, code-phase biases are defined, which can be likened to the difference of the DCM pseudorange specific and carrier-phase-specific clocks. In its implementation, the satellite WL biases are estimated as daily constants, while the station WL biases are estimated epoch per epoch, with a constraint on the overall mean. Similar to the DCM, the full system is defined by a subset of arbitrary integer ambiguities and the remaining WL–NL ambiguities are fixed in bootstrapping integer search algorithms. In practice, for both formulations found above, these arbitrary ambiguity data are constrained using the IF pseudoranges, so the ambiguous satellite and station carrier-phase clocks are somewhat consistent with the pseudorange-specific clocks.

Still other parameterizations than the above given ones are possible as well, like, for example, the common or distinct clock formulations of [25.102, 116]. As all methods provide intrinsically the same external information, one can establish their one-to-one transformations thus showing how the different methods can be mixed between networks and users [25.11]. Instead of the two IF clocks and one WL phase bias, for instance, as used by the DCM and IRC approaches, one can also base the required satellite parameters on the pseudorange-specific IF clock and two between satellite differenced NL–WL network ambiguities or two NL–WL uncalibrated carrier-phase delays [25.110–112].

In the uncalibrated phase delay (UPD) approach of [25.111] and [25.117], the station biases are eliminated through single differencing and the single-differenced UPDs are estimated modulo 1 (WL–NL) cycle. Similar to [25.115], the WL single-difference (SD)-UPDs are estimated as daily constants while the NL SD-UPDs are estimated as piecewise linear polynomials over specified intervals. These network-level products are computed for all satellite combinations, an appropriate selection of which must be matched in the PPP algorithms to provide SD constraints to undifferenced ambiguities.

Another PPP phase ambiguity fixing method, introduced by [25.118], uses DD carrier-phase ambiguities from network DD processing, which are reintroduced into the PPP algorithms as the condition equations of the new undifferenced observations. As it uses information from the global network solution, this method does not require a reparameterization of the observation equations and is very well suited for efficient back-substitution of global results into single station solutions.

In all the above models, PPP solutions require additional network-level products, such as decoupled clocks, pseudorange or carrier-phase biases or UPD/SD-UPDs, to isolate the integer value of ambiguities. However, further development is required to accelerate integer ambiguity resolution and reduce the period of initial convergence, the main operational issue in PPP. From pseudorange initialization at the half-meter to meter level, to the centimeter accuracy attained once all parameters have reached their optimal state, convergence may take 15 min and even longer, depending on receiver-specific pseudorange noise and on the local tracking environment (multipath, ionosphere, antenna dynamics, etc.).

Significant improvements (or even elimination) of PPP convergence is possible when external a priori ionospheric information can be provided. However, this requires precise knowledge of ionospheric delay variations, typically interpolated from local or regional

networks. Furthermore, one has to abandon the ionosphere-free combination approach and instead work with observables that are still sensitive to the ionospheric delays. This can be done, as demonstrated in [25.116, 119–122], by working with the four original measurement types of (25.15) without creating any linear combinations explicitly, or alternatively, as shown in [25.123], by replacing the ionosphere-free Melbourne–Wübbena combination by its carrier

phase and pseudorange constituent parts. Note that such PPP algorithms with fixed atmospheric delays become equivalent to real-time kinematic (RTK), provided that a proper weighting of the observables (undifferenced, DD, or combined) is used [25.11].

Improved results were also shown when using observations made on more frequencies and/or when using more GNSS satellites [25.82, 124–127]. For more details on ambiguity resolution, see Chap. 23.

25.4 Implementations

The availability of global GNSS precise orbits and clocks from various sources has provided the opportunity to develop and implement PPP-based services for positioning and navigation. Post-processed PPP services for both static and kinematic positioning are proving to be particularly useful and efficient for reference frame densification. They have been adopted by several countries as an efficient way to supplement and reduce the expensive infrastructure of dense networks of geodetic monuments traditionally used to provide access to national geodetic reference frames. Internet-based post-processed PPP services are now offered by several institutions to fulfill that function (Chaps. 35 and 36). Many real-time (RT) PPP based positioning and navigation services have also emerged in recent years. RT PPP services are usually more costly to operate and tend to be offered commercially to specialized market segments such as agriculture or land and marine natural resources exploration and exploitation.

It may be useful here to distinguish between online positioning (and navigation) services based on PPP and those based on the differential approach. Although they may appear the same to users who need only to submit GNSS data from one station, they differ fundamentally in their implementation. Differential-based services such as the US National Geodetic Survey Opus and the GeoScience Australia AUSPOS need data from several stations to form the double differences required by their DD processing algorithm. This additional data is normally obtained from their national continuous operating reference stations (CORS) networks as well as from the IGS global network.

Reliance on data from one or more base stations in addition to the one provided by the users has both advantages and disadvantages. On the plus side, the differential services offer more robust cycle-slip detection and repair as well as simpler carrier-phase ambiguity fixing as long as a sufficient number of nearby stations are available. Failure to meet that condition quickly reduces the area of applicability of the differential tech-

nique. While the PPP method can be used globally with almost uniform performance, the differential approach is better suited to regional or continental applications. The remainder of this section will only cover PPP-based services. Irrespective of the specific method, positioning and their users benefit from the continuing standardization for exchange formats for GNSS observation data, the derived products and, to some extent, the resulting solutions (Annex A).

25.4.1 Post-Processed Solutions

Post-processed PPP services are usually more precise than their real-time counterparts and tend to be used for applications requiring accuracy and stability. Although not quite as precise as differential positioning over medium or short baselines (e.g., < 1000 km), post-processed PPP is rapidly being adopted in several regions to establish geodetic control (Chap. 36). This is especially true in remote areas where geodetic control monuments are sparsely distributed or nonexistent. Because of the stability of some of the GNSS orbits and clock products used, post-processed PPP is now providing long-term station velocity estimates for geodynamic applications that compare favorably with those obtained with the differential technique.

Typically for post-processed PPP applications, users submit GNSS data from a single station to the service via the Internet and receive, normally by e-mail and within minutes, an estimated position along with ancillary information. Depending on the service, various formats, such as RINEX, can be used for the GNSS data submission. Very little standardization currently exists for the dissemination of post-processed PPP results. Once received on the host server, GNSS data is processed according to the PPP method using GNSS orbits and clock corrections computed by the service provider or obtained from a third party such as the IGS. In addition to the orbits and clock corrections, PPP services may also call upon other specialized web services for additional corrections such as the ocean tide loading

corrections, troposphere delay parameters and receiver and satellite antenna PCVs.

In addition to the online services, PPP processing is now offered by several GNSS equipment manufacturers within their suite of post-processing software. The computations may be performed within a specialized PPP module or the GNSS data sent to an existing online service. In either case, results are seamlessly integrated into reports and other functions offered by those packages.

25.4.2 Real-Time Solutions

In contrast to the post-processed PPP services that rely on users sending GNSS observation data to a central server, most real-time applications require that GNSS orbit and clock corrections be sent in real time to the point of data collection. PPP position estimation is then performed according to PPP algorithms inside the GNSS receivers or on a colocated computer. Transmission of real-time corrections for PPP is usually done over the Internet using transport protocols developed for that purpose, such as NTRIP (networked transport of Radio Technical Commission for Maritime Services (RTCM) via Internet protocol [25.128]). To ensure that RT PPP services are available even in regions without access to high-speed Internet, some providers are also distributing corrections using geostationary communication satellites, which greatly increase service costs. Most real-time PPP services are currently offered for a cost by commercial enterprises. Costs of those services usually vary with the region as well as with the accuracy required.

Like post-processed PPP services, little standardization currently exists for RT PPP services as most are based on proprietary data formats and customized GNSS end-user equipment. Normally offered by providers of high-precision navigation services in niche markets, GNSS user equipment manufacturers or joint partnerships, RT PPP services tend to be closed-access and fully integrated services, where providers support end-to-end solutions, from computing real-time orbits and clock correction to embedding specialized software in end-user GNSS equipment. This could change in the coming years with the advent of new

standardized formats applicable to RT PPP (Annex A) and Internet-based, free RT PPP corrections such as those of the IGS real-time service (Chap. 33). Open, real-time correction services, although useful for many applications, will require that algorithms in end-user applications be consistent with the models and conventions used to compute the correction streams.

25.4.3 PPP Positioning Services

Listing existing PPP based positioning services is problematic, since such a list can rapidly become outdated or incomplete. Nonetheless, a few examples of post-processed services often referenced in the literature are described in Table 25.3. The listed services all provide static or kinematic processing, use the RINEX observation format and output ITRF estimated coordinates. They, however, provide their results in nonstandard, service-specific output. Many of those services can be accessed and compared at Internet portals such as the University of New Brunswick’s Precise Point Positioning Software Centre [25.129]. Other PPP services may exist that are specific to a given country, region or application.

Although each service has its own user interface, they should all provide comparable position estimates for a specific data set. A thorough assessment of the various services is, however, beyond the scope of this publication. For the interested reader, several papers comparing various post-processed and RT PPP services are available. Having several PPP services that provide independent position estimates also creates some redundancy and increases confidence that a PPP solution can be obtained whenever a particular service is unavailable or suspected of not providing reliable position estimates. However, caution is advised whenever comparing or integrating results from different services to ensure that estimated positions are for the same location, at the same epoch and in the same reference frame. Many services extract information directly from the observation file that is critical to identify the reference point of the position estimates, such as the receiver antenna type and antenna height. Processing reports should be closely examined to ensure that position estimates provided by the various services are indeed compatible.

Table 25.3 Post-processed PPP services

Service	URL	Provider
APPS	http://apps.gdgps.net/	Jet Propulsion Laboratory JPL (National Aeronautics and Space Administration NASA)
CSRS-PPP	http://webapp.geod.nrcan.gc.ca/geod/tools-outils/ppp.php	Natural Resources Canada
GAPS	http://gaps.gge.unb.ca/	University of New Brunswick
MAGIC GNSS	http://magicgnss.gmv.com/	GMV
Trimble RTX	http://www.trimblrtx.com/	Trimble

25.5 Examples

To illustrate the performance offered by the precise point positioning concept, example results of GPS/GLONASS-based PPP solutions are presented in this section. The PPP software of Natural Resources Canada (NRCAN; [25.130] and Table 25.3) has been used, along with 24 h data sets from 17 globally distributed IGS stations observed during the period of Jan. 1, 2012–Feb. 9, 2013. About half of the 17 stations are IGS reference frame stations, that is, a subset of those stations used to align IGS daily solutions to the current ITRF. To reduce computation time and also decrease possible correlation, 24 h PPP solutions using 5 min observation sampling were estimated at five-day intervals.

The NRCAN PPP incorporates all the modeling effects described in Sect. 25.2, including ocean loading, higher-order ionospheric corrections, polar tides and a proper handling of eclipsing satellites for both GPS and GLONASS. However, the atmospheric and hydrological loading effects have not been applied here. Furthermore, only GMF and GPT have been used for tropospheric mapping and a priori ZTD_h, respectively, since GMF/GPT should give slightly better repeatability than the more rigorous VMF1 ones, when no atmospheric loading is applied (Sect. 25.2.1). Also, no ambiguity fixing (Sect. 25.3.4) has been employed. Although only the NRCAN PPP software has been used here, other recent PPP implementations should give similar results.

25.5.1 Static PPP Solutions

Table 25.4 shows the comparison of static PPP positioning solutions with respect to IGS08 reference coordinates using final IGS GPS orbits/clock products as well as European Space Agency (ESA) GPS/GLONASS products. Note that Table 25.4 includes both mean offsets and repeatability of the daily static PPP solutions (one position solution for each 24 h interval), which are affected by real or apparent nonlinear station displacements during this 13-month period. Consequently, the RMS values (the last three columns of Table 25.4) are believed to be a good estimate of the accuracy of static PPP. Here one can notice that the addition of GLONASS has only slightly improved the RMS of the ESA GPS-only PPP solutions. This is likely due to small (real or apparent) systematic effects, common to both GPS and GLONASS PPP solutions. Nevertheless, even the accuracy of the ESA GLONASS-only PPP is quite impressive, considering that it is based on only 24 satellites and that the GLONASS orbit/clock modeling is still being improved.

As expected, the IGS final orbit/clock combinations yielded the smallest RMS of the mean offsets in Table 25.4, though the ESA GLONASS-only PPP RMSs were only slightly larger than the rest. The repeatability of IGS (GPS-only) and ESA GLONASS+GPS PPP solutions are the best. The GLONASS addition improved the longitude repeatability in particular, namely from 3.0 mm down to 2.7 mm.

Ambiguity fixing (Chap. 23), which was not been applied here, would also improve longitude repeatability to a level comparable to what is seen in latitude. The mean offset RMSs, on the other hand, are not expected to be improved by fixing ambiguities, as they are mainly due to station and orbit/clock systematic effects, which are not affected or reduced by ambiguity fixing. However, the initial PPP solution convergence, typically of about 15 min or longer for GPS-only PPP, may be significantly shortened when ambiguities are fixed (Sect. 25.3.4). This is not applicable here, since the 24 h PPP solutions are already fully converged, albeit to nonintegers.

Note that Table 25.4 includes some remote stations with suboptimal performance (e.g., OH12, or the GLONASS HOB2 and DARW data), which have degraded repeatability and mean offset RMS. For most stations the repeatability and RMS are significantly better than those shown in Table 25.4 as can be seen from results previously shown in Fig. 25.3.

25.5.2 Kinematic PPP Solutions

In kinematic mode, independent PPP positions are estimated at each observation epoch, typically every 1–30 s, depending on application and user dynamics. When observation intervals are shorter than the satellite clock sampling, clock interpolation is necessary. Due to satellite clock instability, only clocks at 30 s or lower sampling interval can be reliably interpolated at the cm precision level. IGS and most IGS AC clock solutions currently use 30 s sampling but higher-rate clock products may be offered by individual analysis centers for specific applications [25.131].

To demonstrate kinematic PPP performance, even though under rather ideal conditions, the 17-station static dataset discussed before has also been reprocessed in kinematic mode, where an independent position is solved at each observation (and satellite clock) epoch. Since backward smoothing (substitution) was used for all kinematic PPP solutions, the statistics reflect the solution quality after ambiguity convergence. The resulting kinematic PPP epoch repeatability for each of the 17 stations with the final IGS (GPS-only)

Table 25.4 Repeatability and root mean square (RMS) of the static daily PPP mean offsets (in mm) with respect the IGS08 positions at 17 globally distributed IGS stations, between Jan. 1, 2012 and Feb. 9, 2013, obtained with the final IGS GPS and ESA AC GLONASS/GPS orbits/clocks. Columns dN, dE, and dH refer to the north (latitude), east (longitude) and height (up) component of the position

Static PPP AC orbits/clocks	σ			RMS of means			RMS		
	dN (mm)	dE (mm)	dH (mm)	dN (mm)	dE (mm)	dH (mm)	dN (mm)	dE (mm)	dH (mm)
igs (GPS)	2.2	3.2	6.6	1.7	2.2	3.6	2.8	3.9	7.5
esa (GPS)	2.3	3.0	6.6	2.9	2.3	3.7	3.7	3.8	7.6
esa (GLONASS)	2.4	3.3	7.7	3.0	3.0	4.2	3.8	4.5	8.8
esa (GLONASS+GPS)	2.2	2.7	6.6	2.8	2.5	3.7	3.6	3.7	7.6

and ESA (both GPS and GLONASS) orbits/clocks are shown in Table 25.5.

As expected, the ESA GLONASS+GPS kinematic PPP performed the best, better than GPS-only PPP using the IGS or ESA orbits/clocks. The latitude and longitude of the GLONASS+GPS PPP solutions have sub-cm repeatability at most stations, while the height repeatability is about twice as large (i. e., 2 cm or less at most stations). Note that the RMS differences with respect to IGS08 (not shown here) are about the same as the static PPP RMS, which are much smaller than the kinematic PPP repeatability. Consequently, kinematic PPP repeatability can be seen to represent post-processed kinematic PPP accuracy, achievable under ideal observing conditions (static).

The GLONASS-only kinematic PPP solutions in Table 25.5 performed much worse than the GPS-only ones, with repeatability often exceeding 5 cm. This

should be expected, as there are only 24 GLONASS satellites, yielding weaker geometry and less robust epoch solutions than the 32 GPS satellites. Furthermore, GLONASS orbits/clocks are still less precise and less robust than the GPS ones. Nevertheless, adding GLONASS data already improves kinematic PPP precision or accuracy, sometimes quite significantly. Even for several remote stations (e.g., DARW and HOB2), which had rather large sigmas for GLONASS-only PPP repeatability (Table 25.5), the addition of GLONASS data improved in most cases the GPS-only PPP, sometimes quite significantly. In a real dynamic environment, users are cautioned that kinematic PPP precision can be considerably worse, in particular when operating in real time, which cannot take the advantage of post processing and backward smoothing and is impacted by additional errors due to latencies of real-time clock solutions.

Table 25.5 Repeatability (σ [cm]) of kinematic PPP solutions at 17 IGS stations with IGS and ESA orbit clock products using GPS-only (G), GLONASS-only (R) and GPS+GLONASS (GR) observations (Jan. 1, 2012–Feb. 9, 2013)

Station	North				East				Up			
	IGS G	ESA G	ESA R	ESA RG	IGS G	ESA G	ESA R	ESA RG	IGS G	ESA G	ESA R	ESA RG
CHUR	1.9	1.2	1.8	1.0	2.9	1.4	2.3	0.9	3.8	2.7	3.3	1.5
VALD	0.8	0.9	2.9	1.1	1.0	1.1	4.2	0.8	2.0	2.0	10.9	1.5
WHIT	1.4	1.2	9.2	0.7	1.4	1.4	3.5	0.7	2.6	2.5	14.4	1.4
GOLD	0.8	7.4	71.9	0.7	1.0	16.2	211.9	0.7	2.6	12.9	213.3	1.8
HOFN	1.5	1.3	1.8	0.6	1.1	1.1	3.2	0.5	2.7	2.4	4.5	1.4
ONSA	0.8	0.9	0.9	0.6	0.7	0.7	1.1	0.5	2.2	1.9	2.6	1.4
WTZR	1.0	1.1	1.4	0.7	1.0	1.7	4.2	0.7	2.3	3.3	5.7	1.6
OB2	1.0	1.1	20.7	0.9	1.1	1.2	44.5	1.0	2.9	2.5	169.2	2.0
DARW	1.0	1.1	804.3	1.0	1.3	1.4	981.0	1.5	3.1	3.8	4452	3.0
CHTI	0.8	0.9	1.3	0.6	0.9	1.0	2.1	0.7	2.3	2.1	3.3	1.6
MAW1	2.3	4.3	5.4	0.7	2.6	5.8	8.4	0.6	4.7	4.0	9.6	1.4
HARB	1.1	1.0	1.8	0.8	1.0	1.4	3.1	1.0	2.6	2.7	4.2	2.1
GUUG	1.5	1.7	8.0	1.3	1.8	2.0	14.4	1.5	4.7	5.7	16.6	3.9
GMSD	0.9	1.0	1.5	0.7	1.0	1.1	2.8	0.8	2.8	2.7	5.0	2.1
KOKV	1.8	1.5	18.0	0.7	1.5	1.9	23.7	1.0	3.4	9.0	14.8	2.8
CONZ	0.8	1.0	4.6	0.6	1.0	1.4	7.6	0.9	2.6	2.5	11.4	1.8
OHI2	1.0	1.0	0.9	0.7	1.1	1.1	1.0	0.7	2.2	2.2	2.2	1.6
RMS	1.27	2.34	196	0.81	1.44	4.36	243.78	0.90	3.01	4.75	1082	2.05

25.5.3 Tropospheric Zenith Path Delay

Another parameter estimate available in a PPP solution is the wet tropospheric zenith path delay ZTD_w , which when added to the a priori hydrostatic ZTD_h , yields the total ZTD. When the wet and dry ZTD and the corresponding mapping functions are properly separated (Sect. 25.2.1), ZTD_w can be used to infer the atmospheric precipitable water content (Chap. 38), which can be assimilated into a NWM, although the total ZTD is usually preferred.

Figure 25.4 gives an example of total ZTD PPP solutions at the former IGS station CHAT obtained in 2009 (with an earlier version of the NRCan PPP software), utilizing the IGS Final GPS orbits/clocks. The PPP also estimated stochastic ZTD gradient solutions (Sect. 25.2.1), which are rather small and for brevity are not shown here. The PPP solutions (at 5 min sampling) are compared to the IGS total ZTD products, which are also using 5 min sampling and IGS orbit/clock PPP, but are generated with a different software (GIPSY/OASIS [25.132]). Figure 25.4 also compares the ZTD of the CODE AC, GFZ AC and JPL AC global solutions with the IGS ZTD products. One can see a fairly good agreement for all ZTD solutions, which use the 5 min sampling (IGS, JPL, PPP). The 0.5 h and 2 h samplings, used respectively by GFZ and CODE, are not as responsive to the rapid changes of the total ZTD in the second half of the week (see the ZTD scale on the right). They have likely caused the fairly large (COD-IGS) and (GFZ-IGS) ZTD differences.

The same comparisons seen in Fig. 25.4 have also been done at 33 globally distributed IGS reference sta-

tions resulting in a RMS of 2.8 mm for the (PPP-IGS) ZTD differences. ZTD solutions based on the current PPP software version and the latest IGS orbit/clock and ZTD solutions should give an even better agreement. It is important to realize that the ZTD and height PPP parameters are weakly correlated, namely up to 20% of height errors (either real or apparent) may be mapped into ZTD estimates [25.133]. This is why it is essential that for precise ZTD solutions all the models of Sect. 25.2 be properly taken into account.

25.5.4 Station Clock Solutions

Time and frequency transfer applications of GNSS will be addressed specifically in Chap. 41. This section shows the level of precision at which station clock parameters are recovered using PPP. For this purpose, eight IGS stations (AMC2, BRUX, IENG, NRC1, PTBB, SPT0, USN3, WAB2) located at time and frequency laboratories were processed in three successive 24 h PPP solutions using the NRCan-PPP software in static mode and applying backward smoothing (substitution) to eliminate the initial convergence period. IGS final satellite orbit and clock products were used. The IGS clock products also include stations clock estimates that are consistent with those of satellites, all clocks being referenced to the IGS timescale [25.134].

Figure 25.5 shows the difference of PPP station clock estimates with respect to the IGS Final solutions, where the level of agreement is a few 100 ps (equivalent to 3 cm) or less for the best performing stations, and the solution boundary discontinuities are typical of the systematic effect of pseudorange errors averaged over the 24 h solution interval.

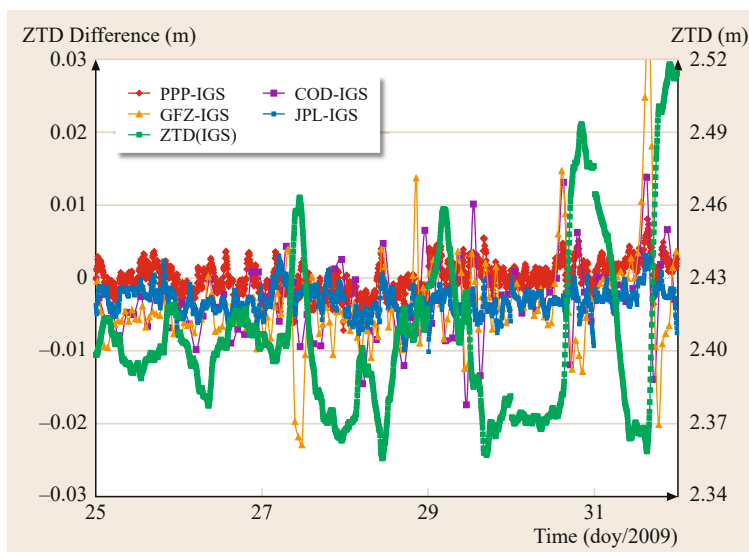


Fig. 25.4 Total ZTD differences from PPP with IGS final orbits/clocks (PPP-IGS) along with Center for Orbit Determination in Europe (CODE), Deutsches GeoForschungsZentrum Potsdam (GFZ) and JPL analysis center solutions with respect to IGS total ZTD at the former IGS station CHAT. The RMS of PPP-IGS ZTD differences are 2.3 mm for station CHAT during this week-long period and 2.8 mm for 33 globally distributed IGS reference stations (IGS, PPP and JPL used 5 min ZTD solution sampling; GFZ and CODE AC used 0.5 h and 2 h ZTD solution intervals, respectively) (after [25.65])

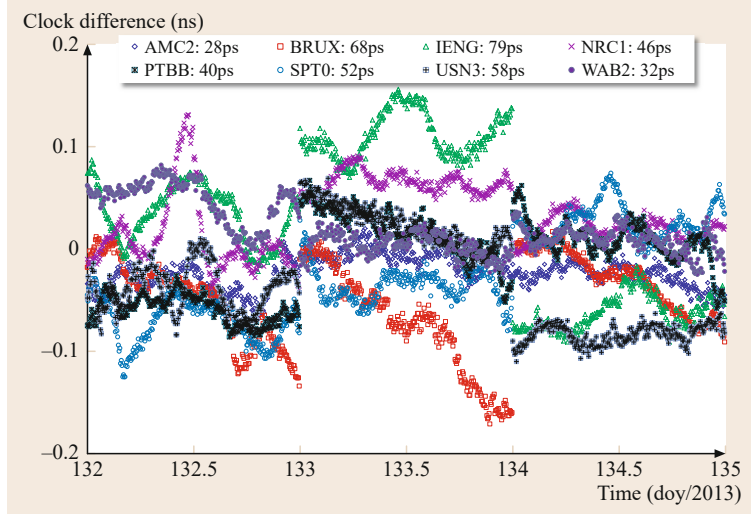


Fig. 25.5 Daily PPP clock solution differences with respect to IGS final solutions for eight IGS stations located at time and frequency laboratories for three consecutive days. RMS differences for each station are below 200 ps

25.6 Discussion

The dual-frequency PPP concept, with related methodology and modeling for static and kinematic PPP solutions with respective precisions of a few mm and one cm, has been reviewed and discussed. Such PPP solutions include position estimates that are directly in the reference frame of the input orbits/clocks, but also facilitate consistent recovery of ZTD solutions at a few-mm level and station clock solutions at the subnanosecond level. Even in single-frequency mode, using ionosphere-free code-plus-phase combinations enables a fairly precise kinematic PPP (navigation) at a few-dm level. This single-frequency PPP can also facilitate ionospheric total electron content (TEC) solutions and monitoring, which compare favorably with the ones based on dual-frequency pseudorange observations.

The PPP solutions can be viewed as an efficient means of realizing the reference frame implied by the fixed orbits/clocks. In fact, they are a station-based back substitution of the global network solutions applied to generate the orbits/clocks products. Consequently, dual-frequency PPP position, ZTD and clock solutions should be as precise as the ones obtained in the corresponding global solutions, provided that the PPP uses consistent models and ambiguity fixing. Ambiguity fixing can improve the PPP positioning precision, particularly the longitude solutions using data spans significantly shorter than 24 h, and it can possibly reduce the time period to initial PPP solution convergence, or even eliminate it, if external, precise ionospheric delays are available.

Even though the above discussions and review pertained mainly to the dual-frequency (L1, L2) PPP with

GPS and GLONASS orbits/clocks, extension to a different frequency pair (e.g., L1, E5) and emerging GNSS constellations, such as BeiDou and Galileo, are straightforward, once all intersystem biases are resolved. Addition of new GNSS signals and satellites will be quite beneficial to PPP solutions, making them more precise and robust, particularly so for kinematic applications. In fact, independent PPP processing of observations from different GNSS systems may provide the redundancy required to facilitate the evaluation of geodetic quantities, as new observation combinations and orbital geometries are exploited to analyze local and long-term systematic effects.

The above discussions have relied heavily on IGS developed conventions (modeling and formats) and IGS orbit/clock products. It is likely that developers and users of PPP will use the IGS orbit/clock solutions, in particular when long time series are to be analyzed with the highest accuracy for consistency with the current IERS standards. It is encouraging to know that the IGS mandate also calls for the provision of orbit/clock solutions for all available GNSSs, as they emerge. Despite their IGS focus, the above discussions should also benefit PPP users of other, for example commercial orbit/clock services, since most of these services benefit from and largely follow the IGS modeling, conventions and developments.

Without IERS conventions and readily available IGS products [25.135] resulting from the significant efforts sustained to develop precise orbit models by many participating organizations, efficient and precise PPP solutions such as those discussed here would not have

been possible. PPP solutions can only be as accurate as implied by the adopted GNSS orbits and clocks!

Acknowledgments. Data and solution products from the International GNSS Service (IGS) have been used in

the preparation of this chapter. Also, the significant help and contributions of the Editors, Oliver Montenbruck and Peter Teunissen, are acknowledged, in particular in regards to the emerging GNSS signals and undifferenced ambiguity resolutions.

References

- 25.1 J.D. Bossler, C.C. Goad, P.L. Bender: Using the global positioning system (GPS) for geodetic positioning, *Bull. Geod.* **54**, 101–114 (1980)
- 25.2 J.F. Zumberge, M.B. Heflin, D.C. Jefferson, M.M. Watkins, F.H. Webb: Precise point positioning for the efficient and robust analysis of GPS data from large networks, *J. Geophys. Res.* **102**, 5005–5017 (1997)
- 25.3 S. Bisnath, Y. Gao: Current state of precise point positioning and future prospects and limitations. In: *Observing Our Changing Earth*, ed. by M.G. Sideris (Springer, Berlin 2009) pp. 615–623
- 25.4 S. Banville, R. Langley: Mitigating the impact of ionospheric cycle slips in GNSS observations, *J. Geodesy* **87**, 179–193 (2013)
- 25.5 R. Schmid, P. Steigenberger, G. Gendt, M. Gendt, M. Rothacher: Generation of a consistent absolute phase center correction model for GPS receiver and satellite antennas, *J. Geodesy* **81**(12), 781–798 (2007)
- 25.6 O. Montenbruck, R. Schmid, F. Mercier, P. Steigenberger, C. Noll, R. Fatkulin, S. Kogure, S. Ganeshan: GNSS satellite geometry and attitude models, *Adv. Space Res.* **56**(6), 1015–1029 (2015)
- 25.7 G.P.S. Directorate: *Navstar GPS Space Segment / Navigation User Segment Interfaces, Interface Specification, IS-GPS-200H*, 24 Sep. 2013 (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo 2013)
- 25.8 J. Kouba: Improved relativistic transformations in GPS, *GPS Solutions* **8**(3), 170–180 (2004)
- 25.9 S. Schaer: Overview of GNSS biases, *Proc. workshop on GNSS biases*, Univ., Bern (2012), (2012)
- 25.10 O. Montenbruck, A. Hauschild, P. Steigenberger: Differential code bias estimation using multi-GNSS observations and global ionosphere maps, *Navigation* **61**(3), 191–201 (2014)
- 25.11 P.J.G. Teunissen, A. Khodabandeh: Review and principles of PPP-RTK methods, *J. Geodesy* **89**(3), 217–240 (2015)
- 25.12 O. Bock, E. Doerflinger: Atmospheric modeling in GPS data analysis for high accuracy positioning, *Phys. Chem. Earth Part A* **26**(6), 373–383 (2001)
- 25.13 G. Petit, B. Luzum: *IERS Conventions (2010)*, IERS Technical Note No. 36 (Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt 2010)
- 25.14 J. Boehm, A. Niell, P. Tregonning, H. Schuh: Global mapping function (GMF): A new empirical mapping function based on numerical weather model data, *Geophys. Res. Lett.* **33**(L07304), 1–4 (2006)
- 25.15 J. Boehm, P. Heinkelmann, H. Schuh: Short note: A global model of pressure and temperature for geodetic applications, *J. Geodesy* **81**(10), 679–683 (2007)
- 25.16 M. Hernández-Pajares, J.M. Juan, J. Sanz, R. Orus, A. García-Rigo, J. Feltens, A. Komjathy, S.C. Schaer, A. Krankowski: The IGS VTEC maps: A reliable source of ionospheric information since 1998, *J. Geodesy* **83**(3/4), 263–275 (2009)
- 25.17 M.M. Hoque, N. Jakowski: Higher order ionospheric effects in precise GNSS positioning, *J. Geodesy* **81**(4), 259–268 (2007)
- 25.18 Z. Altamimi, L.T. Métivier, X. Collilieux: ITRF2008 plate motion model, *J. Geophys. Res.* **117**(B07402), 1–14 (2012)
- 25.19 P.M. Mathews, V. Dehant, J.M. Gipson: Tidal station displacements, *J. Geophys. Res.* **102**(B9), 20469–20477 (1997)
- 25.20 S.A. Melachroinos, R. Biancale, M. Llubes, F. Perosanz, F. Lyard, M. Vergnolle, M.-N. Bouin, F. Masson, J. Nicolas, L. Morel, S. Durand: Ocean tide loading (OTL) displacements from global and local grids: Comparisons to GPS estimates over the shelf of Brittany, France, *J. Geodesy* **82**(6), 357–371 (2008)
- 25.21 M.A. King, Z. Altamimi, J. Boehm, M. Bos, R. Dach, P. Elsegui, F. Fund, M. Hernández-Pajares, D. Lavalée, P.J. Mendes Cerveira, N. Penna, R.E.M. Riva, P. Steigenberger, T. van Dam, L. Vittuari, S. Williams, P. Willis: Improved constraints on models of glacial isostatic adjustment: A review of the contribution of ground-based geodetic observations, *Surveys Geophys.* **31**(5), 465–507 (2010)
- 25.22 T. van Dam, X. Collilieux, J. Wuite, Z. Altamimi, J. Ray: Nontidal ocean loading: Amplitudes and potential effects in GPS height time series, *J. Geodesy* **86**(11), 1043–1057 (2012)
- 25.23 L. Petrov, J.-P. Boy: Study of the atmospheric pressure loading signal in very long baseline interferometry observations, *J. Geophys. Res.* **109**(B03405), 1–14 (2004)
- 25.24 B. Görres, J. Campbell, M. Becker, M. Siemes: Absolute calibration of GPS antennas: Laboratory results and comparison with field and robot techniques, *GPS Solutions* **10**(2), 136–145 (2006)
- 25.25 J.T. Wu, S.C. Wu, G.A. Hajj, W.I. Bertiger, S.M. Lichten: Effects of antenna orientation on GPS carrier-phase, *Man. Geodetica* **18**, 91–98 (1993)

- 25.26 A.Q. Le, C.C.J.M. Tiberius: Phase wind-up effects in precise point positioning with kinematic platforms, Proc. NAVITEC, Noordwijk (2006) pp. 1–8, (ESA, Noordwijk 2006)
- 25.27 P. Steigenberger: Accuracy of Current and Future Satellite Navigation Systems, Habilitation Thesis (Fakultät Bau Geo Umwelt, Technische Universität München, Munich 2015)
- 25.28 W.M. Folkner, J.G. Williams, D.H. Boggs: *The Planetary and Lunar Ephemeris DE421*, Memorandum IOM 343R-08-003 (Jet Propulsion Laboratory, Pasadena 2008)
- 25.29 J.L. Hilton, C.Y. Hohenkerk: A comparison of the high accuracy planetary ephemerides DE421, EPM2008, and INPOP08, Proc. Journées, 2010, “Systèmes de Référence Spatio-Temporels” (JSR2010): New challenges for reference systems and numerical standards in astronomy, Paris, ed. by N. Capitaine (Observatoire de Paris, Paris 2011) pp. 77–80
- 25.30 H.F. Fliegel, K.M. Harrington: Sun/Moon position routines for GPS trajectory calculations, Proc. AIAA/AAS Astrodyn. Conf., Hilton Head Island (1992) pp. 625–631
- 25.31 J.H. Meeus: *Astronomical Algorithms* (Willmann-Bell, Richmond 1991)
- 25.32 O. Montenbruck, E. Gill: *Satellite Orbits – Models, Methods and Applications* (Springer, Berlin 2000)
- 25.33 H.A. Marques, J.F.G. Monico, M. Aquino: RINEX_H0: Second- and third-order ionospheric corrections for RINEX observation files, GPS Solutions **15**(3), 305–314 (2011)
- 25.34 N. Jakowski, F. Porsch, G. Mayer: Ionosphere-induced-ray-path bending effects in precise satellite positioning systems, Z. Satell. Position. Navig. Kommun. **SPN 1/94**, 6–13 (1994)
- 25.35 Z. Wang, Y. Wu, K. Zhang, Y. Meng: Triple-frequency method for high-order ionospheric refractive error modelling in GPS modernization, J. Glob. Position. Syst. **1**(9), 291–295 (2005)
- 25.36 J. Saastamoinen: Atmospheric correction for the troposphere and stratosphere in radio ranging of satellites. In: *The Use of Artificial Satellites for Geodesy*, Geophysical Monograph, Vol. 15, ed. by S.W. Henriksen, A. Mancini, B.H. Chovitz (AGU, Washington 1972) pp. 247–251
- 25.37 J.L. Davis, T.A. Herring, I.I. Shapiro, A.E.E. Rogers, G. Elgered: Geodesy by radio interferometry: Effects of atmospheric modeling errors on estimates of baseline length, Radio Sci. **20**(6), 1593–1607 (1985)
- 25.38 G. Chen, T.A. Herring: Effects of atmospheric azimuthal asymmetry on the analysis of space geodetic data, J. Geophys. Res. **102**(B9), 20489–20502 (1997)
- 25.39 D.S. MacMillan, C. Ma: Atmospheric gradients and the VLBI terrestrial and celestial reference frames, Geophys. Res. Lett. **24**(4), 453–456 (1997)
- 25.40 J.W. Marini: Correction of satellite tracking data for an arbitrary tropospheric profile, Radio Sci. **7**(2), 223–231 (1972)
- 25.41 K.M. Lagler, M. Schindelegger, J. Bohm, H. Krasna, T. Nilsson: GPT2: Empirical slant delay model for radio space geodetic techniques, Geophys. Res. Lett. **40**(6), 1069–1073 (2013)
- 25.42 J. Boehm, B. Werl, H. Schuh: Troposphere mapping functions for GPS and very long baseline interferometry from European centre for medium-range weather forecasts operational analysis data, J. Geophys. Res. **111**(B02406), 1–9 (2006)
- 25.43 L. Urquhart, F.G. Nievinski, M.C. Santos: Assessment of troposphere mapping functions using three dimensional raytracing, GPS Solutions **18**(3), 345–354 (2014)
- 25.44 GGOS Atmosphere – *Atmosphere Delays* (Vienna Univ. Technology, Vienna 2014) <http://ggosatm.hg.tuwien.ac.at/delay.html>
- 25.45 UNB Vienna Mapping Function Service (Univ. New-Brunswick, Fredericton 2015) <http://unb-vmf1.gge.unb.ca/>
- 25.46 J. Kouba: Implementation and testing of the grid-der Vienna mapping function 1 (VMF1), J. Geodesy **82**, 193–205 (2008)
- 25.47 J. Kouba: Testing of global pressure/temperature (GPT) model and global mapping function (GMF) in GPS analyses, J. Geodesy **83**(3), 199–208 (2009)
- 25.48 R. Schmid: Upcoming switch to IGS08/igs08.atx – Details on igs08.atx (IGS Mail 6355; International GNSS Service, 7 Mar. 2011) <http://igs08.jpl.nasa.gov/pipermail/igs08mail/2011/006347.html>
- 25.49 G. Wübbena, M. Schmitz, G. Boettcher, C. Schumann: Absolute GNSS antenna calibration with a robot: Repeatability of phase variations, calibration of GLONASS and determination of carrier-to-noise pattern, Proc. IGS Workshop, Darmstadt (ESA/ESOC, Darmstadt 2006) pp. 1–12
- 25.50 RINEX – The Receiver Independent Exchange Format – Version 3.03 14 July 2015 (IGS RINEX WG and RTCM-SC104, 2015)
- 25.51 L. Scott: Why do GNSS systems use circular polarization antennas?, Inside GNSS **2**(2), 30–33 (2007)
- 25.52 M.L. Psiaki, S. Mohiuddin: Modeling, analysis, and simulation of GPS carrier phase for spacecraft relative navigation, J. Guid. Contr. Dyn. **30**(6), 1628–1639 (2007)
- 25.53 Y.E. Bar-Sever: A new module for GPS yaw attitude control, Proc. IGS Workshop – Special Topics and New Directions, Potsdam (Geoforschungszentrum, Potsdam 1996) pp. 128–140
- 25.54 J. Kouba: A simplified yaw-attitude model for eclipsing GPS satellites, GPS Solutions **13**(1), 1–12 (2009)
- 25.55 F. Dilssner, R. Springer, G. Gienger, J. Dow: The GLONASS-M satellite yaw-attitude model, Adv. Space Res. **47**(1), 160–171 (2010)
- 25.56 S. Banville, H. Tang: Antenna rotation and its effects on kinematic precise point positioning, Proc. ION GNSS 2010, Portland (ION, Virginia 2010) pp. 2545–2552
- 25.57 D.D. McCarthy: *IERS Conventions (1989)*, IERS Technical Note No. 3 (Observatoire de Paris, Paris 1989)
- 25.58 J.M. Wahr: The forced nutation of an elliptical, rotating, elastic, and ocean less Earth, Geophys.

- 25.59 J. R. Astron. Soc. **64**, 705–727 (1981)
G. Jentzsch: Earth tides and ocean tidal loading. In: *Tidal Phenomena*, ed. by H. Wilhelm, H.G.W. Wenzel Zürich (Springer, Berlin 1997) pp. 145–171
- 25.60 H. Dragert, T.S. James, A. Lambert: Ocean loading corrections for continuous GPS: A case study at the Canadian coastal site Holberg, *Geophys. Res. Lett.* **27**(14), 2045–2048 (2000)
- 25.61 H.G. Scherneck: A parameterized solid earth tide model and ocean tide loading effects for global geodetic baseline measurements, *Geophys. J. Int.* **106**, 677–694 (1991)
- 25.62 Online ocean tide loading computation service (Chalmers University) <http://holt.oso.chalmers.se/loading/>
- 25.63 O. Montenbruck, A. Hauschild: Code biases in multi-GNSS point positioning, *Proc. ION ITM 2013*, San Diego (ION, Virginia 2013) pp. 616–628
- 25.64 O. Montenbruck, P. Steigenberger, R. Khachikyan, G. Weber, R.B. Langley, L. Mervart, U. Hugentobler: IGS-MGEX: Preparing the ground for multi-constellation GNSS science, *Inside GNSS* **9**(1), 42–49 (2014)
- 25.65 J. Kouba: *A Guide to Using International GNSS Service (IGS) Products* (IGS, Pasadena 2015), <http://kb.igs.org/>
- 25.66 International GNSS Service: Analysis center information <ftp://igsceb.jpl.nasa.gov/igsceb/center/analysis/>
- 25.67 R.R. Hatch: The synergism of GPS code and carrier measurements, *Proc. Third Int. Geodetic Symp. Satellite Doppler Positioning*, Las Cruces (Physical Science Laboratory, Las Cruces 1982) pp. 1213–1232
- 25.68 O. Øvstedal: Absolute positioning with single-frequency GPS receivers, *GPS Solutions* **5**(4), 33–44 (2002)
- 25.69 A.Q. Le, C. Tiberius: Single-frequency precise point positioning with optimal filtering, *GPS Solutions* **11**(1), 61–69 (2007)
- 25.70 R.J.P. van Bree, C.C.J.M. Tiberius: Real-time single-frequency precise point positioning: Accuracy assessment, *GPS Solutions* **16**(2), 259–266 (2012)
- 25.71 A. Tetewsky, J. Ross, A. Soltz, N. Vaughn, J. Anzperger, Ch. O'Brien, D. Graham, D. Craig, J. Lozow: Making sense of inter-signal corrections – Accounting for GPS satellite calibration parameters in legacy and modernized ionosphere correction algorithms, *Inside GNSS* **4**(4), 37–48 (2009)
- 25.72 P. Héroux, J. Kouba: GPS precise point positioning with a difference, *Geomatics'95*, Ottawa (1995) pp. 1–11
- 25.73 T.P. Yunck: Coping with the atmosphere and ionosphere in precise satellite and ground positioning. In: *Environmental Effects on Spacecraft Positioning and Trajectories*, ed. by A.V. Jones (AGU, Washington 1992), Chap. 1, pp. 1–16
- 25.74 H. Van Der Marel, P. De-Bakker: Single versus dual-frequency precise point positioning – What are the tradeoffs between using L1-only and L1+L2 for PPP?, *Inside GNSS* **7**(4), 30–35 (2012)
- 25.75 S. Choy, K. Zhang, D. Silcock: An evaluation of various ionospheric error mitigation methods used in single frequency PPP, *J. Glob. Position. Syst.* **7**(1), 62–71 (2008)
- 25.76 T. Schüller, H. Diessongo, Y. Poku-Gyamfi: Precise ionosphere-free single-frequency GNSS positioning, *GPS Solutions* **15**(2), 139–147 (2011)
- 25.77 H.T. Diessongo, H. Bock, T. Schüller, S. Junker, A. Kiroe: Exploiting the Galileo E5 wideband signal for improved single-frequency precise positioning, *Inside GNSS* **7**(5), 64–73 (2012)
- 25.78 K. Chen, Y. Gao: Real-time precise point positioning using single frequency data, *Proc. ION GNSS 2005*, Long Beach (2005), pp. 1514–1523
- 25.79 S. Banville, R.B. Langley: Cycle-slip correction for single-frequency PPP, *Proc. ION GNSS 2012*, Nashville (ION, Virginia 2012) pp. 3753–3761
- 25.80 C. Cai, Y. Gao: Precise point positioning using combined GPS and GLONASS observations, *J. Glob. Position. Syst.* **6**(1), 13–22 (2007)
- 25.81 L. Wanninger, S. Wallstab-Freitag: Combined processing of GPS, GLONASS, and SBAS code phase and carrier phase measurements, *Proc. ION GNSS 2007*, Fort Worth (ION, Virginia 2007) pp. 866–875
- 25.82 C. Cai, Y. Gao: Modeling and assessment of combined GPS/GLONASS precise point positioning, *GPS Solutions* **17**(4), 223–236 (2013)
- 25.83 T. Melgard, E. Vigen, O. Orpen: Advantages of combined GPS and GLONASS PPP – Experiences based on G2, a new service from Fugro, *Proc. 13th IAIN World Congress*, Stockholm (IAIN, London 2009) pp. 1–7
- 25.84 S. Choy, S. Zhang, F. Lahaye, P. Héroux: A comparison between GPS-only and combined GPS+GLONASS precise point positioning, *J. Spatial Sci.* **58**(2), 169–190 (2013)
- 25.85 L. Wanninger: Carrier-phase inter-frequency biases of GLONASS receivers, *J. Geodesy* **86**(2), 139–148 (2012)
- 25.86 J.M. Sleewaegen, A. Simsky, W. Boon, F. de Wilde, T. Willems: Demystifying GLONASS inter-frequency carrier-phase biases, *Inside GNSS* **7**(3), 57–61 (2012)
- 25.87 M. Becker, P. Zeimet, E. Schönmann: Anechoic chamber calibrations of phase center variations for new and existing GNSS signals and potential impacts in IGS processing, *Proc. IGS Workshop*, Newcastle (IGS, Pasadena 2010) pp. 1–44
- 25.88 P. Steigenberger, U. Hugentobler, S. Loyer, F. Perosanz, L. Prange, R. Dach, M. Uhlemann, G. Gendt, O. Montenbruck: Galileo orbit and clock quality of the IGS multi-GNSS experiment, *Adv. Space Res.* **55**(1), 269–281 (2015)
- 25.89 Y. Lou, Y. Liu, C. Shi, X. Yao, F. Zheng: Precise orbit determination of BeiDou constellation based on BETS and MGEX network, *Sci. Rep.* **4**(4692), 1–10 (2014)
- 25.90 P. Steigenberger, A. Hauschild, O. Montenbruck, C. Rodríguez-Solano, U. Hugentobler: Orbit and clock determination of QZS-1 based on the CONGO network, *Navigation* **60**(1), 31–40 (2013)

- 25.91 L. Prange, E. Orliac, R. Dach, D. Arnold, G. Beutler, S. Schaer, A. Jäggi: CODE's multi-GNSS orbit and clock solution, Proc. EGU General Assembly, Vienna (EGU, Munich 2015) p. 11494
- 25.92 J. Tegner, O. Øvstedal, E. Vigen: Precise orbit determination and point positioning using GPS, Glonass, Galileo and BeiDou, *J. Geod. Sci.* **4**(1), 65–73 (2014)
- 25.93 D. Odijk, P.J.G. Teunissen: Characterization of between-receiver GPS-Galileo inter-system biases and their effect on mixed ambiguity resolution, *GPS Solutions* **17**(4), 521–533 (2013)
- 25.94 D. Odijk, P.J.G. Teunissen: Estimation of differential inter-system biases between the overlapping frequencies of GPS, Galileo, BeiDou and QZSS, Proc. 4th Int. Coll. Scientific and Fundamental Aspects of the Galileo Programme, Prague 2013 (ESA, Noordwijk 2013) pp. 1–8
- 25.95 A. Dalla Torre, A. Caporali: An analysis of intersystem biases for multi-GNSS positioning, *GPS Solutions* **19**(2), 297–307 (2015)
- 25.96 J. Paziewski, P. Wielgosz: Accounting for Galileo-GPS inter-system biases in precise satellite positioning, *J. Geodesy* **89**(1), 81–93 (2015)
- 25.97 H. Cui, G. Tang, S. Hu, B. Song, H. Liu, J. Sun, P. Zhang, C. Li, M. Ge, C. Han: Multi-GNSS processing combining GPS, GLONASS, BDS and GALILEO observations, Proc. CSNC, Nanjing, Vol. III (2014), ed. by J. Sun, W. Jiao, H. Wu, M. Lu (Springer, Berlin 2014) pp. 121–132
- 25.98 X. Li, X. Zhang, X. Ren, M. Fritsche, J. Wickert, H. Schuh: Precise positioning with current multi-constellation global navigation satellite systems: GPS, GLONASS, Galileo and BeiDou, *Sci. Rep.* **5**(8328), 1–14 (2015)
- 25.99 X. Li, M. Ge, X. Dai, X. Ren, M. Fritsche, J. Wickert, H. Schuh: Accuracy and reliability of multi-GNSS real-time precise positioning: GPS, GLONASS, BeiDou, and Galileo, *J. Geodesy* **89**(6), 607–635 (2015)
- 25.100 P.J.G. Teunissen, R. Odolinski, D. Odijk: Instantaneous BeiDou+GPS RTK positioning with high cut-off elevation angles, *J. Geodesy* **88**(4), 335–350 (2014)
- 25.101 J. Tegner, O. Øvstedal: Triple carrier precise point positioning (PPP) using GPS L5, *Survey Rev.* **46**(337), 288–297 (2014)
- 25.102 P.J.G. Teunissen, D. Odijk, B. Zhang: PPP-RTK: Results of CORS network-based PPP with integer ambiguity resolution, *J. Aeronaut. Astronaut. Aviat., Ser. A* **42**(4), 223–230 (2010)
- 25.103 E. Schönemann: Analysis of GNSS Raw Observations in PPP Solutions, Ph.D. Thesis (TU Darmstadt, Darmstadt 2013)
- 25.104 H. Chen, W. Jiang, M. Ge, J. Wickert, H. Schuh: Efficient high-rate satellite clock estimation for PPP ambiguity resolution using carrier-ranges, *Sensors* **14**(12), 22300–22312 (2014)
- 25.105 O. Montenbruck, U. Hugentobler, R. Dach, P. Steigenberger, A. Hauschild: Apparent clock variations of the block IIF-1 (SVN-62) GPS satellite, *GPS Solutions* **16**(3), 303–313 (2012)
- 25.106 G. Wübbena, M. Schmitz, A. Bagg: PPP-RTK: Precise point positioning using state-space representation in RTK networks, Proc. ION GNSS 2005, Long Beach (ION, Virginia 2005) pp. 13–16
- 25.107 D. Laurichesse, F. Mercier: Integer ambiguity resolution on undifferenced GPS phase measurements and its application to PPP, Proc. ION GNSS 2007, Fort Worth (ION, Virginia 2007) pp. 839–848
- 25.108 L. Mervart, Z. Lukes, C. Rocken, T. Iwabuchi: Precise point positioning with ambiguity resolution in real-time, Proc. ION GNSS 2008, Savannah (ION, Virginia 2008) pp. 397–405
- 25.109 P. Collins: Isolating and estimating undifferenced GPS integer ambiguities, Proc. ION NTM 2008, San Diego (ION, Virginia 2008) pp. 720–732
- 25.110 M. Ge, G. Gendt, M. Rotacher, C. Shi, J. Liu: Resolution of GPS carrier-phase ambiguities in precise point positioning (PPP) with daily observations, *J. Geodesy* **82**(7), 389–399 (2008)
- 25.111 W. Bertiger, S. Dessai, B. Haines, N. Harvey, A.W. Moore, S. Owen, P. Weiss: Single receiver phase ambiguity resolution with GPS data, *J. Geodesy* **84**(5), 3337 (2010)
- 25.112 J. Geng, F.N. Teferle, X. Meng, A.H. Dodson: Towards PPP-RTK: Ambiguity resolution in real-time precise point positioning, *Adv. Space Res.* **47**(10), 1664–1673 (2011)
- 25.113 A. Lannes, J.L. Prieur: Calibration of the clock-phase biases of GNSS networks: The closure-ambiguity approach, *J. Geodesy* **87**(8), 709–731 (2013)
- 25.114 P. Collins, S. Bisnath, F. Lahaye, P. Héroux: Undifferenced GPS ambiguity resolution using the decoupled clock model and ambiguity datum fixing, *Navigation* **57**(2), 123–135 (2010)
- 25.115 D. Laurichesse, F. Mercier, J.P. Berthias, P. Broca, L. Cerri: Integer ambiguity resolution on undifferenced GPS phase measurements and its application to PPP and satellite precise orbit determination, *Navigation* **56**(2), 135–149 (2009)
- 25.116 B. Zhang, P.J.G. Teunissen, D. Odijk: A novel undifferenced PPP-RTK concept, *J. Navigation* **64**(S1), 180–191 (2011)
- 25.117 J. Geng, C. Shi, M. Ge, A.H. Dodson, Y. Lou, Q. Zhao, J. Liu: Improving the estimation of fractional-cycle biases for ambiguity resolution in precise point positioning, *J. Geodesy* **86**(8), 579–589 (2013)
- 25.118 G. Blewitt: Fixed point theorems of GPS carrier-phase ambiguity resolution and their application to massive network processing: Ambizap, *J. Geophys. Res.* **113**(B12410), 1–12 (2008)
- 25.119 D. Odijk, P.J.G. Teunissen, B. Zhang: Single-frequency integer ambiguity resolution enabled GPS precise point positioning, *J. Survey Eng.* **138**(4), 193–202 (2012)
- 25.120 L. Mervart, C. Rocken, T. Twabuchi, Z. Lukes, M. Kanzaki: Precise point positioning with fast ambiguity resolution prerequisites, algorithms and performance, Proc. ION GNSS 2013, Nashville (ION, Virginia 2013) pp. 1176–1185
- 25.121 X. Li, M. Ge, H. Zhang, J. Wickert: A method for improving uncalibrated phase delay estimation and

- ambiguity fixing in real-time precise point positioning, *J. Geodesy* **87**(5), 405–416 (2013)
- 25.122 S. Banville, P. Collins, P. Héroux, P. Tétreault, P.F. Lahaye: Precise cooperative positioning: A case study in Canada, *Proc. ION GNSS 2014*, Tampa (ION, Virginia 2014) pp. 2503–2511
- 25.123 P. Collins, F. Lahaye, S. Bisnath: External ionospheric constraints for improved PPP-AR initialisation and a generalised local augmentation concept, *Proc. ION GNSS 2012*, Nashville (ION, Virginia 2012) pp. 3055–3065
- 25.124 J. Geng, Y. Bock: Triple-frequency GPS precise point positioning with rapid ambiguity resolution, *J. Geodesy* **87**(5), 449–460 (2013)
- 25.125 L. Pan, C. Cai, R. Santerre, J. Zhu: Combined GPS/GLONASS precise point positioning with fixed GPS ambiguities, *Sensors* **14**, 17530–17547 (2014)
- 25.126 D. Odijk, B. Zhang, P.J.G. Teunissen: Multi-GNSS PPP and PPP-RTK: Some GPS+BDS results in Australia, *Proc. CSNC (2015) Vol. II*, Xi'an, ed. by J. Sun, J. Liu, S. Fan, X. Lu (Springer, Berlin 2015) pp. 613–623
- 25.127 L. Qu, Q. Zhao, J. Guo, G. Wang, X. Guo, Q. Zhang, K. Jiang, L. Luo: BDS/GNSS real-time kinematic precise point positioning with un-differenced ambiguity resolution, *Proc. CSNC, Vol. III (2015)*, Xi'an, ed. by J. Sun, J. Liu, S. Fan, X. Lu (Springer, Berlin 2015) pp. 13–29
- 25.128 G. Weber, D. Dettmering, H. Gebhard, R. Kalafus: Networked transport of RTCM via internet protocol (Ntrip) – IP-streaming for real-time GNSS applications, *Proc. ION GPS 2005*, Long Beach (ION, Virginia 2005) pp. 2243–2247
- 25.129 The Precise Point Positioning Center (Univ. New-Brunswick, Fredericton 2015), <http://gge.unb.ca/Resources/PPP/Purpose.html>
- 25.130 P. Héroux, J. Kouba: GPS precise point positioning using IGS orbit products, *Phys. Chem. Earth (A)* **26**(6–8), 573–578 (2001)
- 25.131 H. Bock, R. Dach, A. Jäggi, G. Beutler: High-rate GPS clock corrections from CODE: Support of 1 Hz applications, *J. Geodesy* **83**(11), 1083–1094 (2009)
- 25.132 S.H. Byun, Y.E. Bar-Sever: A new type of troposphere zenith path delay product of the international GNSS service, *J. Geodesy* **83**(3/4), 1–7 (2009)
- 25.133 G. Gendt: IGS combination of tropospheric estimates – Experience from pilot experiment, *Proc. Anal. Center Workshop (1998) Darmstadt*, ed. by J.M. Dow, J. Kouba, T. Springer (IGS, Pasadena 1998) pp. 205–216
- 25.134 K. Senior, P. Koppang, D. Matsakis, J. Ray: Developing an IGS time scale, *Proc. IEEE Freq. Contr. Symp. 2001*, Seattle (IEEE, Washington 2001) pp. 211–218
- 25.135 J. Dow, R.E. Neilan, G. Gendt: The International GPS Service (IGS): Celebrating the 10th anniversary and looking to the next decade, *Adv. Space Res.* **36**, 320–326 (2005)

Differential P

26. Differential Positioning

Dennis Odijk, Lambert Wanninger

Part E | 26.1

This chapter describes the concepts of differential global navigation satellite system (DGNSS) positioning focusing on practical details given that the fundamental concepts have been covered in prior chapters. The chapter starts with a review of the general concepts of DGNSS, including a quantitative discussion on the biases in DGNSS measurements. The next section focusses on code-based DGNSS positioning, presenting an overview of DGNSS services as well as a brief discussion on the format and latency of DGNSS corrections. A significant part of this chapter is devoted to carrier-phase dominated DGNSS, or real-time kinematic (RTK) positioning. Besides a theoretical consideration that includes the Russian Global Navigation Satellite System (GLONASS) and multi-GNSS RTK, the section provides examples of RTK positioning performance that are obtained in practice. The last section details on network RTK, which is an extension of the standard RTK technique to cover longer distances.

26.1 Differential GNSS: Concepts	753
26.1.1 Differential GNSS Observation Equations	753
26.1.2 Differential GNSS Biases	754
26.2 Differential Navigation Services	760
26.2.1 DGNSS Implementations	760
26.2.2 DGNSS Services	761
26.2.3 Data Communication: RTCM Message	762
26.2.4 Latency of DGNSS Corrections	762
26.3 Real-Time Kinematic Positioning	763
26.3.1 Double-Differenced Positioning Model	763
26.3.2 Carrier-Phase-Based Positioning Methods	764
26.3.3 GLONASS RTK Positioning	766
26.3.4 Multi-GNSS RTK Positioning	768
26.3.5 RTK Positioning Examples	770
26.4 Network RTK	774
26.4.1 From RTK to Network RTK	774
26.4.2 Data Processing Methods for Network RTK	774
26.4.3 Network RTK Correction Models	776
26.4.4 Refined Virtual Reference Stations	777
26.4.5 From Network RTK to PPP-RTK	778
References	778

26.1 Differential GNSS: Concepts

This section describes and compares the concepts of differential GNSS (DGNSS) positioning. We will address the code (pseudorange)-based DGNSS positioning techniques as well as the more precise carrier-phase-dominated DGNSS positioning techniques. Pseudorange DGNSS remains an important method for obtaining meter level positions, whereas real-time kinematic (RTK) carrier-phase DGNSS receiver systems and correction services are essential tools in surveying and many other fields. Furthermore, network-based approaches to deliver RTK carrier-phase DGNSS services have more recently gained importance.

Before discussing the differential positioning techniques, Sect. 26.1.1 reviews the global navigation satel-

lite system (GNSS) code and phase observation equations, as they form the basis of the positioning models.

26.1.1 Differential GNSS Observation Equations

Reference is made to the basic (linearized) observation equations as presented in Chaps. 19 and 21, but now with the satellite position vector still assumed as unknown in the observation equations (in Chap. 21 it was assumed that the satellite positions are known, such that they disappear from the linearized positioning model). If, similar to Chap. 21, it is assumed that there are two receivers, a rover (or remote) receiver denoted by r and a reference (pivot or base) receiver denoted

by 1, that both track data of satellite s that corresponds to GNSS constellation S at the same frequency j , the between-receiver differenced (linearized) observation equations for pseudorange (code) and carrier-phase read

$$\begin{aligned}\Delta p_{1r,j}^s &= -\mathbf{e}_r^{s\top} \Delta \mathbf{r}_{1r} + \mathbf{e}_{1r}^{s\top} (\Delta \mathbf{r}^s - \Delta \mathbf{r}_1) \\ &\quad + T_{1r}^s + \mu_j^s I_{1r}^s + c [dt_{1r} + d_{1r,j}^s \\ &\quad + \Delta d_{1r,j}^s] + \mathbf{e}_{1r,j}^s, \\ \Delta \varphi_{1r,j}^s &= -\mathbf{e}_r^{s\top} \Delta \mathbf{r}_{1r} + \mathbf{e}_{1r}^{s\top} (\Delta \mathbf{r}^s - \Delta \mathbf{r}_1) \\ &\quad + T_{1r}^s - \mu_j^s I_{1r}^s + c [dt_{1r} + \delta_{1r,j}^s \\ &\quad + \Delta \delta_{1r,j}^s] + \lambda_j^s N_{1r,j}^s + \varepsilon_{1r,j}^s.\end{aligned}\quad (26.1)$$

The following notation is used:

$\Delta p_{r,j}^s$	Observed-minus-computed code
$\Delta \varphi_{r,j}^s$	Observed-minus-computed phase
\mathbf{e}_r^s	Line-of-sight vector of unit length
$\Delta \mathbf{r}_1$	Incremental pivot receiver position
$\Delta \mathbf{r}_{1r}$	Incremental relative receiver position
$\Delta \mathbf{r}^s$	Incremental satellite position
T_{1r}^s	Differential tropospheric delay
μ_j^s	Ionospheric coefficient for j -th freq.
I_{1r}^s	Differential ionospheric delay
c	Velocity of light
dt_{1r}	Differential receiver clock
$d_{1r,j}^s$	Differential receiver code bias
$\delta_{1r,j}^s$	Differential receiver phase bias
$\Delta d_{1r,j}^s$	Differential code interchannel bias
$\Delta \delta_{1r,j}^s$	Differential phase interchannel bias
λ_j^s	Wavelength for frequency j
$N_{1r,j}^s$	Differential carrier-phase ambiguity
$\varepsilon_{1r,j}^s$	Differential code noise
$\varepsilon_{1r,j}^s$	Differential phase noise.

Note that $(\cdot)^\top$ denotes the transpose of a matrix or vector. The between-receiver differencing has eliminated the satellite clock as well as satellite hardware biases from the observation equations. Compared to the undifferenced observation equations, all other parameters are changed to their between-receiver differential counterparts. Similar to Chap. 21, the between-receiver differenced observables and parameters are denoted by $(\cdot)_{1r} = (\cdot)_r - (\cdot)_1$. Note that besides the (incremental) satellite position vector (i. e., $\Delta \mathbf{r}^s$), also the (incremental) position vector of the reference receiver 1 (i. e., $\Delta \mathbf{r}_1$) is maintained in the differential observation equations (instead of assuming it as known). The reason for this is as to evaluate the impact of errors in either one of them on the differential observation equations (Knowing the satellite positions as well as the pivot receiver position implies that $\Delta \mathbf{r}^s = 0$, as well as $\Delta \mathbf{r}_1 = 0$

(Sect. 26.1.2). If the position of the reference receiver is only approximately known, using DGNSS precise *position differences* between reference and rover (also referred to as *baseline coordinates*), denoted by \mathbf{r}_{1r} , are estimated.

An alternative approach for DGNSS is that instead of forming differences of observations between rover and pivot receivers, *pseudorange corrections* and/or *phase-range corrections* (Chap. 21) are formed based on the data of the pivot receiver. These *differential corrections* are then transmitted to the rover receiver to correct its observations.

Finally, we remark that the differential *interchannel* (or *interfrequency*) biases in (26.1), that is, $\Delta d_{1r,j}^s$ and $\Delta \delta_{1r,j}^s$, only appear in case of GLONASS (frequency division multiple access (FDMA)) observations. These interchannel biases (ICBs) exist as each GLONASS signal transmits on its own frequency channel (Chap. 8). In case of code division multiple access (CDMA) signals these interchannel biases are absent.

26.1.2 Differential GNSS Biases

The main motivation for the development of differential positioning techniques was the presence of *Selective Availability* (SA) on Global Positioning System (GPS) signals. SA was implemented to deliberately degrade the GPS positioning performance after it was discovered that single point positioning (SPP) based on the civil C/A-code performed better than originally expected [26.1]. SA implied that the GPS broadcast ephemerides were manipulated and the satellite clock stability was degraded (*dithering*) by the U.S. government [26.2]. This had as consequence that the accuracy of C/A-code-based SPP (i. e., the GPS standard positioning service) was only about 100 m. SA has been turned off since 2 May 2000, resulting in a SPP accuracy of about 10 m, an improvement of a factor 10. As an illustration, Fig. 26.1 shows the position errors obtained using SPP for an Australian international GNSS service (IGS) station (YAR1) during the first 8.3 h of 2 May 2000. From the graphs it can be directly seen that after SA was turned off at about 04:07 UTC, the position accuracy improved tremendously.

Although the main motivation for the development of differential techniques was to eliminate the effects of SA, DGNSS has remained a very important positioning method after SA was turned off, because of its elimination and reduction of biases. Besides the elimination of the satellite clock and hardware biases, biases due to orbit and atmospheric are significantly reduced, based on their spatial correlation. The remainder of this section provides a quantitative assessment of the different types of DGNSS biases.

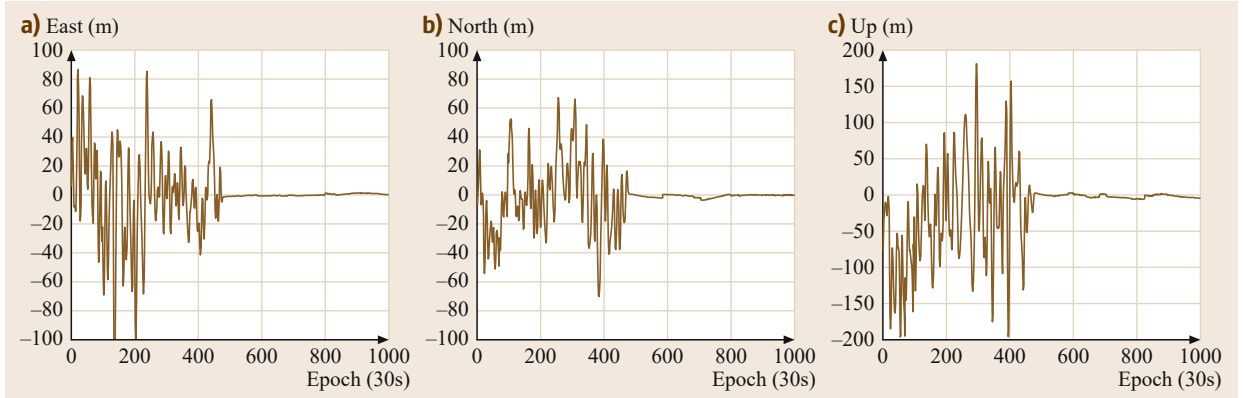


Fig. 26.1a–c East (a), north (b), and up (c) position errors of station YAR1 (Western Australia) obtained with single point positioning during the first 8.3 h Coordinated Universal Time (UTC) of 2 May 2000. The effect of turning off *Selective Availability* at about 04:07 UTC (epoch 480 in the graphs) is clearly visible

Satellite and Pivot Receiver Position Biases

The satellite position as well as the position of the pivot receiver are multiplied by the differential line-of-sight vector \mathbf{e}_{1r}^s in the observation equations (26.1). The effect of a bias in either the satellite position or pivot receiver position can be evaluated as follows. Recall (21.84) in Chap. 21, which gives a rule-of-thumb for the impact of a bias in the satellite position on the relative baseline

$$|\mathbf{e}_{1r}^{s\top} \Delta \mathbf{r}^s| \leq \frac{\|\mathbf{r}_{1r}\|}{\|\mathbf{r}^s - \mathbf{r}_r\|} \|\Delta \mathbf{r}^s\|. \quad (26.2)$$

Here $\|\mathbf{r}_{1r}\|$ denotes the baseline length between reference and rover receivers, $\|\mathbf{r}^s - \mathbf{r}_r\|$ is the distance between receiver and satellite and $\|\Delta \mathbf{r}^s\|$ the size of the bias in the satellite position. The impact of a bias in the position of the pivot receiver, denoted by $\|\Delta \mathbf{r}_1\|$, is upper bounded in a similar manner

$$|\mathbf{e}_{1r}^{s\top} \Delta \mathbf{r}_1| \leq \frac{\|\mathbf{r}_{1r}\|}{\|\mathbf{r}^s - \mathbf{r}_r\|} \|\Delta \mathbf{r}_1\|. \quad (26.3)$$

The accuracy of broadcast GPS ephemerides is about 1 m. Using the above rule-of-thumb, this has an effect of $1000 \text{ mm} / 20\,000 \text{ km} = 0.05 \text{ ppm}$ of the baseline length. For a baseline of 1000 km, the bias is then 5 cm. The accuracy of precise GPS ephemerides is better than 5 cm (Chap. 33), resulting in a bias of $50 \text{ mm} / 20\,000 \text{ km} = 0.0025 \text{ ppm}$ of the baseline length. This 5 cm bias only has an effect of 2.5 mm for a 1000 km baseline. Figure 26.2a shows the bias in the baseline as a function of the bias in either satellite position or pivot receiver position for four baseline lengths. From this figure, it can be inferred that biases in satellite position and/or pivot receiver position get drastically reduced in the between-receiver differenced observation equations.

The fact that the coordinates of the pivot or reference receiver need not be accurate for the differential solution to be precise means that DGNSS can be accomplished by setting the pivot receiver coordinates to any reasonably valid position solution. This fact allows for *moving reference station DGNSS* which has applications in navigation, specifically in vehicle-to-vehicle relative navigation and also in terrestrial or space formation flying. In such applications, the absolute location of the rover is not required, rather the relative location of two moving vehicles is of interest.

Differential Ionospheric Biases

As the ionospheric delays are spatially correlated, the between-receiver differenced ionospheric bias is much less than its absolute counterparts. To get insight into the size of this differential ionospheric bias, we make use of a simple mapping of the (slant) ionospheric delays to the vertical ionospheric delays at a representative height of the ionosphere above the Earth's surface (Chap. 19)

$$I_r^s = \frac{1}{\cos z_r^{s'}} I_v, \quad (26.4)$$

$$z_r^{s'} = \arcsin \left(\frac{R_\oplus}{R_\oplus + h_{\text{ion}}} \sin z_r^s \right).$$

Here I_v denotes the vertical ionospheric delay at height h_{ion} of the ionospheric layer, z_r^s the zenith angle at the receiver, $z_r^{s'}$ the zenith angle at the ionospheric layer, and R_\oplus the Earth's radius. Figure 26.2b depicts the geometry of this single-layer ionosphere in relation to the between-receiver difference.

The vertical ionospheric delay at the ionospheric point of receiver r is, in principle, different from the vertical delay of receiver 1, due to the existence of hor-

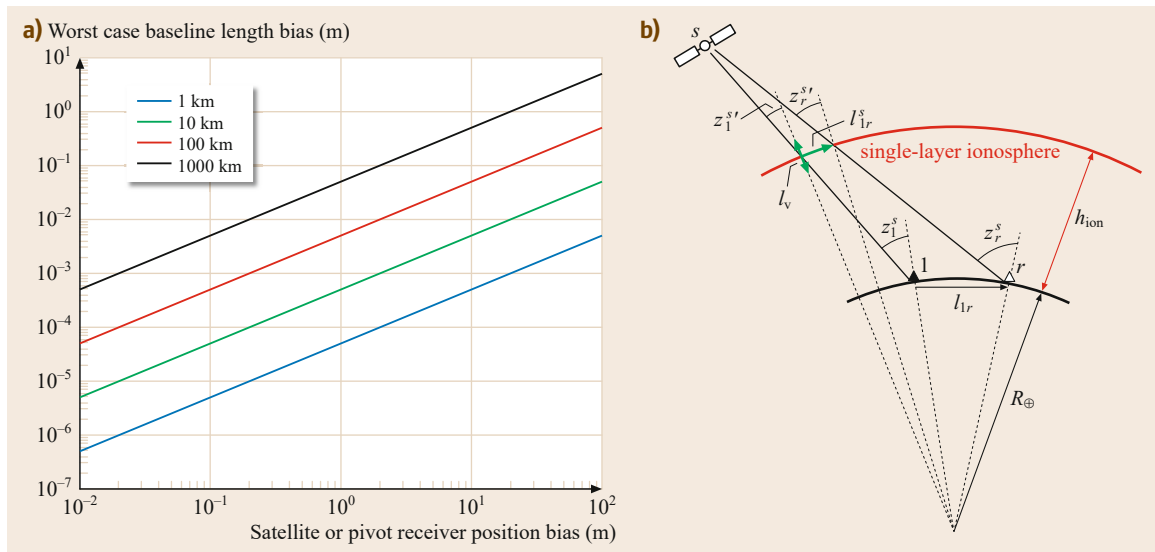


Fig. 26.2 (a) Baseline length bias as a function of a bias in satellite position or pivot receiver position for four baseline lengths. (b) Geometry of a single-layer ionosphere model in relation to the between-receiver difference for stations 1 and r

horizontal gradients in the ionosphere. If it is assumed that the vertical delay of receiver r is equal to

$$I_v + \frac{\partial I_v}{\partial l} l_{1r}^s,$$

where $\partial I_v / \partial l$ denotes the horizontal gradient and l_{1r}^s the baseline length at the ionospheric single layer (Fig. 26.2b), the between-receiver ionospheric bias can be decomposed as follows

$$l_{1r}^s = - \left(\frac{1}{\cos z_1^{s'}} - \frac{1}{\cos z_r^{s'}} \right) I_v + \frac{l_{1r}^s}{\cos z_r^{s'}} \frac{\partial I_v}{\partial l}. \quad (26.5)$$

Figure 26.3 now plots for different baselines, varying from 10 to 400 km, the functions $(1/\cos z_1^{s'} - 1/\cos z_r^{s'})$ as well as $l_{1r}^s/\cos z_r^{s'}$ as a function of the zenith angle at receiver 1 (z_1^s). For these graphs, the height of the ionospheric layer was set to $h_{\text{ion}} = 350$ km. Besides that the graphs show that these functions become larger with increasing baseline length, it can be seen that for all baselines they reach a maximum around a zenith angle of $z_1^s \approx 75^\circ$. This also means that their combination in (26.5) and thus the differential ionospheric bias is maximized at this zenith angle. Furthermore, from Fig. 26.3 follows that at this zenith angle $(1/\cos z_1^{s'} - 1/\cos z_r^{s'})$ is about 0.6 ppm of the baseline length l_{1r} , whereas $l_{1r}^s/\cos z_r^{s'}$ is about 1.5 ppm of the baseline length l_{1r} .

Using these values for a zenith angle of 75° (this corresponds to an elevation of 15°), the size of the differential ionospheric bias is assessed based on worst-case choices for the vertical delay I_v and the horizontal gradient $\partial I_v / \partial l$. These choices, which are given in Table 26.1, apply to levels for the vertical total electron content (VTEC) that can be reached during the day-time in either solar minimum or solar maximum years, as the ionospheric activity is clearly correlated with the progression of the 11-yearly solar cycle. These VTEC values have been converted from total electron content (TEC) unit to the ionospheric delay at the GPS L1 frequency in meters. Difference is made in levels that apply to mid-latitude regions and equatorial regions, as the levels around the equator are normally much higher than at mid-latitudes. Levels for polar regions are not included in the table, as they vary more or less in between the mid-latitude and equatorial levels. The equatorial vertical delay levels in Table 26.1

Table 26.1 Typical worst-case values for vertical ionospheric delay I_v and horizontal ionospheric gradient $\partial I_v / \partial l$ during solar minimum and maximum years for mid-latitude and equatorial regions

	Mid-latitudes	Equatorial region
Solar minimum	$I_v = 3$ m $\frac{\partial I_v}{\partial l} = 2$ ppm	$I_v = 12$ m $\frac{\partial I_v}{\partial l} = 2$ ppm
Solar maximum	$I_v = 6$ m $\frac{\partial I_v}{\partial l} = 10$ ppm	$I_v = 22$ m $\frac{\partial I_v}{\partial l} = 50$ ppm

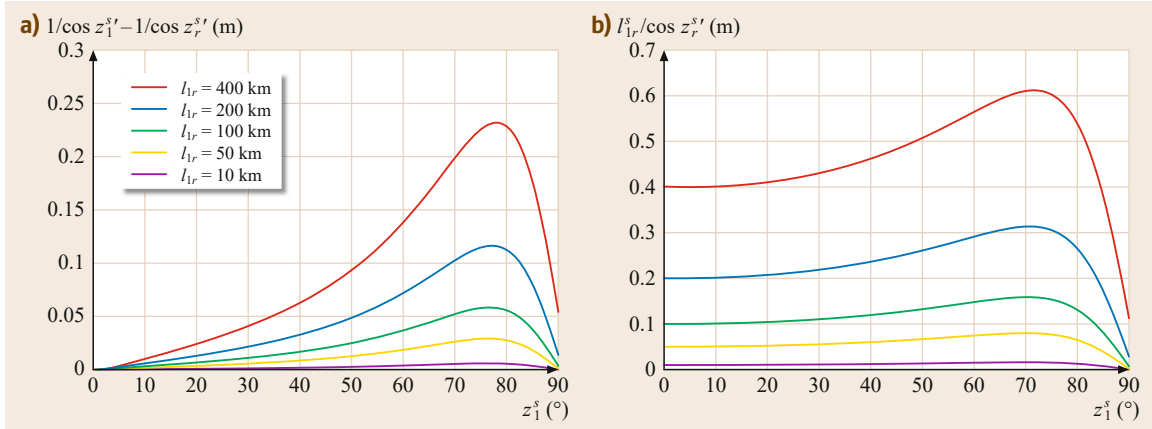


Fig. 26.3 (a) Amplification of $I_v = 1$ m into I_{1r}^s ; (b) amplification of $\frac{\partial I_v}{\partial l} = 1$ ppm into I_{1r}^s , for different baseline lengths

are confirmed by [26.3, 4], whereas the mid-latitude levels are confirmed by, among others, [26.5, 6]. Concerning the horizontal gradients, the normal east–west gradient due to the diurnal cycle of the ionosphere is about 1 ppm. So, during solar minimum years they are not much larger than this; in Table 26.1 we have assumed 2 ppm for both mid-latitude and equatorial regions. Much larger gradients are usually observed close to solar maximum: whereas horizontal gradients at mid-latitudes can range up to 10 ppm [26.7], very large (north–south) gradients of almost 50 ppm have been reported by [26.8, 9] for the equatorial anomaly regions, occurring post sunset.

Using the values of Table 26.1, the differential ionospheric bias is calculated using (26.5) for a zenith angle at the reference receiver of 75° . Table 26.2 presents the absolute value of this differential ionospheric bias, in ppm of the baseline length. From this table, it follows that the size of the differential ionospheric bias varies significantly, in the sense of depending on the location on Earth as well as in what stage of the solar cycle the GNSS measurements are collected. The lowest values are obtained during solar minimum at mid-latitudes (1 ppm), whereas the highest values (62 ppm) apply to solar maximum at equatorial regions.

Depending on the positioning application at hand, these differential ionospheric bias levels put restric-

tions to the maximum baseline length for which they can be neglected. For *carrier-phase-based DGNSS positioning*, requiring the highest positioning accuracy (Sect. 26.3), the differential ionospheric biases are allowed to be only a few centimeters in order to neglect them. For example, if it is required that $|I_{1r}^s| \leq 2.5$ cm; based on Table 26.2 the baseline length may be up to 25 km during solar minimum at mid-latitudes. However, this is only about 0.5 km during solar maximum for baselines that are measured in equatorial regions. For *code-based DGNSS applications* that require a much lower positioning accuracy (Sect. 26.2), the differential ionospheric biases can be ignored for much longer baselines.

Differential Tropospheric Biases

The tropospheric bias is normally decomposed into a *hydrostatic* plus a *wet* component (Chap. 6). The hydrostatic delay, which equals about 90% of the tropospheric bias, can be predicted very well based on the surface’s air pressure. The wet component (accounting for about 10%) is caused by atmospheric water vapor and is more variable and therefore harder to predict. Usually the hydrostatic component is corrected using an a priori model, whereas the wet component is estimated as unknown parameter in the processing after having it *mapped* to local zenith.

We may get an impression of the size of the zenith tropospheric bias by using *Saastamoinen’s* troposphere model [26.10]. Mapping both hydrostatic and wet delays to local zenith, the zenith hydrostatic delay, denoted by T_h^z and zenith wet delay, denoted by T_w^z , can be calculated as

$$\begin{aligned} T_h^z &= Bp \\ T_w^z &= B \left(\frac{1255}{T} + 0.05 \right) e \end{aligned} \quad (26.6)$$

Table 26.2 Worst-case (absolute) differential ionospheric bias $|I_{1r}^s|$ in ppm of the baseline length for mid-latitude and equatorial regions during solar minimum and maximum years

	Mid-latitudes	Equatorial region
Solar minimum	1 ppm	4 ppm
Solar maximum	11 ppm	62 ppm

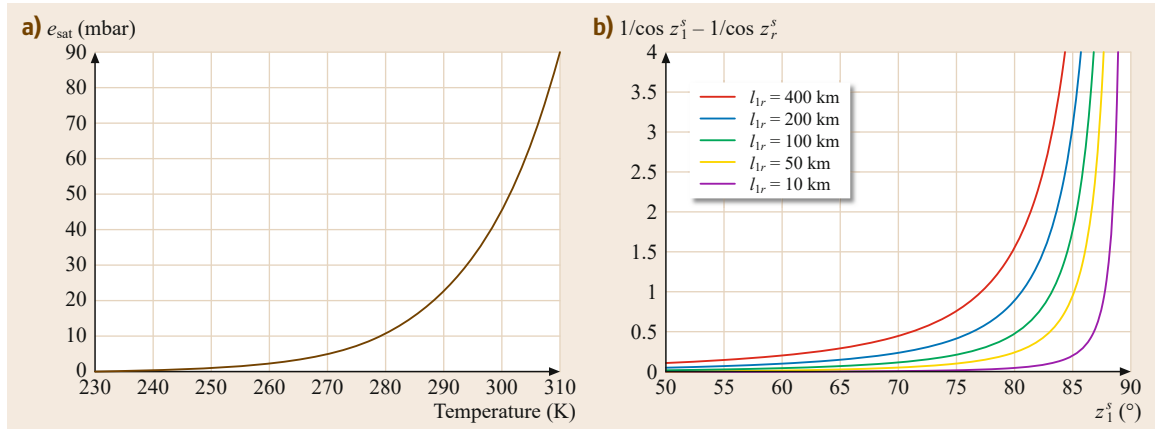


Fig. 26.4 (a) Partial pressure of saturated air as a function of temperature; (b) differential tropospheric mapping function as a function of baseline lengths ranging from 10–400 km

with the constant $B = 0.002277 \text{ m/mbar}$. Here p denotes the air pressure (mbar), T the temperature (K) and e the partial pressure of water vapor (mbar) at the Earth's surface. This partial pressure can be computed as follows [26.11]

$$e = \text{rh} e_{\text{sat}}$$

$$e_{\text{sat}} = \exp \left(a - \frac{b}{T} \right). \quad (26.7)$$

Here rh denotes the relative humidity ($0 \leq \text{rh} \leq 1$) and e_{sat} the partial pressure of *saturated* air, which can be modeled as an exponential function of the (inverse) temperature. This function, for which its constants a and b are taken from [26.12], is plotted in Fig. 26.4a for the temperature ranging between 230 and 310 K.

According to the above functions, Table 26.3 presents the calculated values for the hydrostatic, wet, and total delay in zenith for two types of tropospheric conditions, that is, *cold and dry* as well as *hot and humid*. Under these two extreme atmospheric conditions, from Table 26.3 it follows that the difference in (total) zenith delay can be almost 1 m. Note that the *hot and humid* values of about 2.4 m for the zenith hydrostatic delay and about 0.8 m for the zenith wet delay are maximum values that can be reached [26.13].

To assess the size of the differential tropospheric bias, the reference receiver is assumed to experience

Table 26.3 Examples of extreme atmospheric conditions and zenith hydrostatic, wet and total delays for reference receiver 1

	p (mbar)	T (K)	rh (%)	$T_{1,h}^z$ (m)	$T_{1,w}^z$ (m)	T_1^z (m)
Cold+dry	950	230	10	2.16	0.00	2.16
Hot+humid	1050	310	90	2.39	0.75	3.14

the tropospheric conditions as in Table 26.3. For the troposphere at the rover receiver, it is first assumed that it is at another *height*, as it is well known that both temperature and pressure in the lower part of the atmosphere decreases with altitude. The temperature and pressure at height H (km) can be assessed as follows, based on a constant temperature lapse rate of 6.5 K/km [26.11]

$$T_H = T_0 - 6.5H,$$

$$p_H = p_0 \left(\frac{T_H}{T_0} \right)^{\mu+1} \quad \text{with } \mu \approx 4, \quad (26.8)$$

where T_0 and p_0 denote the temperature and pressure at $H = 0$, respectively. For a (rover) receiver that is 1000 m higher than the receiver in Table 26.3, Table 26.4 presents the tropospheric zenith delays under the two *extreme* conditions, where it is assumed that relative humidities are same. Besides a difference in height, tropospheric biases between receivers may also experience differences due to *horizontal gradients*, which may occur as a result of (cold) weather fronts. Therefore, we also consider a rover receiver that is at the same height as the reference receiver, but experiences a temperature and pressure that are 10 K and 10 mbar, respectively, lower than at the reference receiver. Table 26.5 presents the calculated zenith tropospheric

Table 26.4 Examples of extreme atmospheric conditions and zenith hydrostatic, wet and total delays, for receiver r with a height difference of +1000 m with respect to the receiver in Table 26.3

	p (mbar)	T (K)	rh (%)	$T_{r,h}^z$ (m)	$T_{r,w}^z$ (m)	T_r^z (m)
Cold+dry	823.1	223.5	10	1.87	0.00	1.87
Hot+humid	944.1	303.5	90	2.15	0.50	2.65

Table 26.5 Examples of extreme atmospheric conditions and zenith hydrostatic, wet and total delays, for receiver r at the same height as the receiver in Table 26.3, but temperature and pressure are, respectively, 10 K and 10 mbar lower due to a horizontal gradient in the troposphere (cold weather front)

	p (mbar)	T (K)	rh (%)	$T_{r,h}^z$ (m)	$T_{r,w}^z$ (m)	T_r^z (m)
Cold+dry	940	220	10	2.14	0.00	2.14
Hot+humid	1040	300	90	2.37	0.40	2.77

delays for this rover receiver (again assuming the same relative humidities).

Based on these examples, we can now assess the size of the differential (slant) tropospheric bias between reference and rover receivers. For this purpose, the wet delay is mapped to zenith using a simple $1/\cos z$ mapping function, such that it is possible to decompose the between-receiver tropospheric bias as follows

$$T_{1r}^s = -\left(\frac{1}{\cos z_1^s} - \frac{1}{\cos z_r^s}\right) T_1^z + \frac{1}{\cos z_r^s} (T_r^z - T_1^z). \quad (26.9)$$

Here T_1^z and T_r^z denote the zenith tropospheric delay at receivers 1 and r , respectively. Similar to the differential ionospheric bias, the differential tropospheric bias shows a dependence on the length of the baseline through factor $(1/\cos z_1^s - 1/\cos z_r^s)$. This factor is plotted for several baseline lengths in Fig. 26.4b. In contrast to the ionospheric factor (Fig. 26.3a), this tropospheric factor does not have a maximum for a certain zenith angle; it is rapidly increasing for high zenith angles or low elevations. For the assessment of the differential tropospheric biases, we therefore assume that there is a cut-off zenith angle of 75° , above which the data are not used. For this cut-off angle, Table 26.6 shows the magnitude of the differential tropospheric biases that can be expected for the examples. It is assumed that the baseline length is 100 km. Besides the results for the baseline for which the rover is 1000 m higher than the reference and the baseline that experiences a cold front, results are included for a baseline for which the atmospheric conditions at both receivers are the same. For such a baseline, the differential tropospheric biases are only a result of the fact that both receivers see the satellite at different elevations.

In addition to the total differential tropospheric bias, Table 26.6 also gives the differential zenith wet delays, which give an impression of the amount of remaining bias after application of an a priori troposphere model. From the table follows that the size of the total differential tropospheric biases varies between 0.54 m (identical cold and dry conditions for both receivers)

Table 26.6 Worst-case differential tropospheric bias $|T_{1r}^s|$ in meter for a 100 km baseline and the atmospheric conditions in Tables 26.3–26.5. Between brackets are the differential zenith wet delays $|T_{1r,w}^z|$. The satellite cut-off elevation is 15°

	Identical atmospheric conditions	Height difference of 1000 m	Cold weather front
Cold+dry	0.54 (0.00)	1.67 (0.00)	0.62 (0.00)
Hot+humid	0.79 (0.19)	2.70 (1.16)	2.23 (1.55)

and 2.70 m (hot and humid conditions and a height difference of 1 km between both receivers). This corresponds to 5–27 ppm of the baseline length. After a priori correction, the remaining differential biases are much smaller, although for the *hot and humid* conditions they cannot be neglected for this 100 km baseline and need to be estimated. For the *cold and dry* conditions as well as sufficiently short baselines they can be ignored.

Receiver Noise, Multipath, and Other Biases

Biases that are *not* reduced or cancelled in between-receiver differential GNSS, as they are not spatially correlated, are first of all differential receiver clocks and differential receiver hardware biases. For the carrier-phase observations, this also includes the differential ambiguities. Hence, these biases need to be estimated as unknown parameters.

Biases that remain are *unmodeled* biases, due to for example, ionospheric scintillations, radio interferences, multipath, signal scattering, signal attenuation, and diffraction [26.14]. Of these additional biases, *multipath* biases (Chap. 15) are most likely to dominate. Multipath affects both code and phase measurements, but for phase data the multipath bias is generally 100 times smaller (centimeter level) than for code data (meter level) [26.15]. Multipath is a highly localized phenomenon and cannot be removed by the differential positioning approach; any multipath experienced by the reference receiver will be directly passed to the rover receiver. To mitigate this, the reference station location must be carefully chosen and a multipath mitigating receiver and antenna should be used. The use of multiple reference stations at different locations to average out multipath contributed to the development of *wide-area DGPS services* (WADGPS) [26.16], (Sect. 26.2).

Finally, random *receiver noise* is not cancelled when forming differential GNSS measurements. Although receiver noise depends on the type of make of receiver and antenna [26.17], we find some theoretical lower bounds on the precision of high-end geodetic GPS receivers in [26.14]: 1 dm for code observations

and 0.1 mm for phase observations. In practice, these values are however larger, as unmodeled biases cannot be separated from receiver noise. For GPS phase data, the precision is at the level of a few millimeters, whereas for GPS code data this is a few decimeters

(for high-end receivers, [26.18]). It has been initially demonstrated that the receiver noise of newer signals (e.g., GPS L5) or of new constellations (Galileo, BeiDou navigation satellite system (BDS)) is lower than of the GPS L1 and L2 signals [26.19–22].

26.2 Differential Navigation Services

In this section, code-based DGNSS is discussed. First, Sect. 26.2.1 describes the DGNSS implementations that exist in practice, followed by an overview of DGNSS services in Sect. 26.2.2. Section 26.2.3 focusses on the message format that is used for transmitting the DGNSS corrections to users, whereas Sect. 26.2.4 discusses the issues due to the difference in time between the generation of the DGNSS corrections and the time they are applied by users (i. e., latency).

26.2.1 DGNSS Implementations

The principle of DGNSS is visualized in Fig. 26.5. A GNSS reference receiver that is stationed at a known location tracks data of all satellites in view and determines differential (DGNSS) corrections that are transmitted (in real time) to users (rover receivers) that track GNSS data at a certain distance of the reference station. These users correct their data allowing them to improve their positioning accuracy compared to a SPP solution.

If the corrections are determined based on the data of a single reference station (Fig. 26.5a), the technique is referred to as *local DGNSS* [26.23]. The rover receiver is usually of the single-frequency type (GPS: L1). The maximum distance the rover receiver is allowed to be is about 1000 km, as to have sufficient satellites in common between rover and reference. Because of the distance-dependent biases (atmosphere, orbit), the position accuracy of local *DGPS* is restricted

to 1–10 ppm [26.24]. For a distance of 1000 km, the position accuracy is then about 10 m, which corresponds to the accuracy of SPP, so it does not make sense to cover larger distances. The position accuracy can be improved if DGNSS is based on a whole *network* of reference stations, instead a single reference station (Fig. 26.5b). All reference stations, equipped with dual-frequency or multifrequency receivers at known locations, send their data to a central processing facility which generates DGNSS corrections that are valid for the region the network covers. This is the *wide-area DGNSS* (WADGNSS) technique. WADGNSS has the following advantages over local DGNSS [26.25]:

- A network covers an effective area that is *larger* than based on a single reference station.
- Differential corrections are *consistent* within the area covered by the network, whilst based on different single reference stations discrepancies may show up.
- A network is able to calculate *models for distance-dependent biases* (atmospheric errors, orbits) that are valid over the coverage area, while this is not possible based on a single reference station.

Note that the above advantages also apply to network RTK as compared to single reference RTK (Sect. 26.4). By using a network of reference stations, the coverage area of DGNSS can span the size of the United States or Europe, or even the world. The positioning accuracy

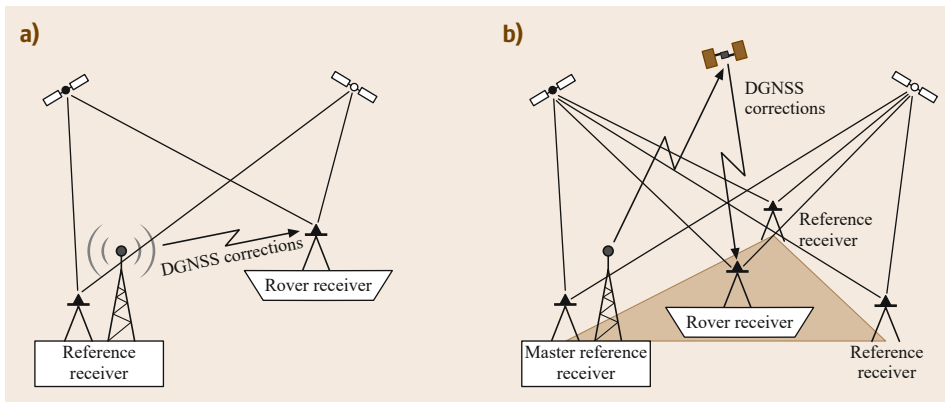


Fig. 26.5 (a) Local DGNSS, based on a single reference station, where the corrections to the rover are sent by a ground-based antenna, vs. **(b)** Wide-area DGNSS, based on a network of reference stations, with the corrections transmitted via a satellite

of (code-only) *WADGPS* has been reported to be within 5 m [26.26]. Higher, even submeter accuracies are feasible for *WADGPS*, when the code data are *smoothed* with carrier-phase data [26.27]. The combined use of code with carrier-phase then makes *WADGNSS* conceptually equivalent to *precise point positioning* (*PPP*) (Chap. 25).

The positioning accuracy of local (single-frequency) DGNSS can be improved by a priori reducing the differential ionospheric biases by means of an ionosphere model of which its coefficients are broadcast in the navigation message. In case of GPS or BDS this is the Klobuchar model [26.28, 29], whereas for Galileo this is the semiempirical NeQuick model, which is based on the algorithm proposed by [26.30]. These models correct for some part of the differential ionospheric biases, but not completely. The *Klobuchar* model claims to remove 50–80% of the magnitude of the ionospheric bias [26.31], while it has been initially demonstrated that the *NeQuick* model performs better than *Klobuchar* [26.32]. Differential tropospheric biases should be a priori corrected for using standard models, for example, the Saastamoinen model [26.10] or the empirical global pressure and temperature (*GPT*) model [26.33].

26.2.2 DGNSS Services

DGNSS: Marine and Terrestrial Applications

By far the largest present application of DGNSS is in the *marine transport* sector. This is primarily because during the 1990s, the coast guards and lighthouse authorities of several maritime nations deployed extensive networks of reference stations providing differential corrections to mitigate the effects of SA on GPS. Almost immediately after the implementation of SA by the U.S. Air Force, the U.S. Federal Aviation Admin-

stration, and the U.S. Coast Guard began to make plans to correct for SA using differential techniques. This resulted in the *U.S. nationwide DGPS* (NDGPS) system that is jointly operated by the U.S. Coast Guard and the U.S. Department of Transportation [26.34]. The NDGPS network (operational since 1999) initially made use of existing marine radio beacons as a method for transmitting differential corrections to users and originally only covered coastal areas. Later on also reference stations were deployed at inland locations, such that the whole U.S. is covered by the NDGPS network. In 2016 it was decided to maintain coverage in maritime and coastal regions only [26.35], due to the growth of other Continuously Operating Reference Station (CORS) networks inland.

DGNSS Service Providers: GBAS or SBAS

On the transmission of the DGNSS corrections, one can distinguish between DGNSS services that broadcast the differential corrections through terrestrial messages, such as the U.S. NDGPS, or systems that broadcast these messages using (geostationary or geosynchronous) satellites. The first type of systems are referred to as *ground-based augmentation systems* (*GBAS*) (Chap. 31), whereas those of the second type are *satellite-based augmentation systems* (*SBAS*) (Chap. 12). The transmission through satellites has the advantage that a larger area can be covered compared to the transmission by means of terrestrial links.

Examples of governmental GBAS- and SBAS-based DGNSS services can be found in Table 26.7, which provides an overview of some (but not all) DGNSS service providers. Although this table only mentions a few countries that provide a nationwide DGNSS service, many countries actually have their governmental DGNSS services, covering both coastal (offshore) and/or inland (onshore).

Table 26.7 Examples of governmental and commercial real-time WADGPS service providers (code-only services)

Provider name	Governmental/Commercial	Coverage area	GBAS/SBAS
NDGPS	Governmental	United States	GBAS
SAPOS EPS	Governmental	Germany	GBAS
AMSA	Governmental	Australia	GBAS
LAAS	Governmental	United States	GBAS
WAAS	Governmental	United States	SBAS
EGNOS	Governmental	Europe	SBAS
MSAS	Governmental	Japan	SBAS
GAGAN	Governmental	India	SBAS
QZSS	Governmental	Japan	SBAS
OmniSTAR VBS (Trimble)	Commercial	Global	SBAS
StarFire (NavCom/J. Deere)	Commercial	Global	SBAS
Starfix.L1 (Fugro)	Commercial	Global	SBAS
Veripos Standard (Veripos)	Commercial	Global	SBAS
Thales DGPS DGRS 610/615	Commercial	Flexible	GBAS

Note that the governmental Wide Area Augmentation System (WAAS), European Geostationary Navigation Overlay Service (EGNOS), Multi-Function Satellite Augmentation System (MSAS), and GPS Aided Geo Augmented Navigation (GAGAN) systems are all of the SBAS type which are discussed in Chap. 12. Although the MSAS and Quasi-Zenith Satellite System (QZSS) are both Japanese systems, they differ in the types of services they offer: whereas MSAS is a WAAS or EGNOS compatible system that provides corrections to augment single-frequency code-based positioning, QZSS provides corrections to augment positioning (including high-accuracy carrier-phase-based positioning) in the densely built Japanese urban regions (Chap. 11). The local area augmentation system (LAAS), which is used for the precision approach and landing of aircraft (Chap. 31), is of the GBAS type as it uses a very high frequency (VHF) radio link for the transmission of DGPS corrections.

In addition to these governmental services, commercial service providers exist that operate on a global scale (Table 26.7). Commercial DGNSS systems are either of the GBAS or the SBAS type. Nowadays, these commercial providers offer not only code-based DGPS services, but also multi-GNSS (GPS+GLONASS+BDS+Galileo+QZSS) code and carrier-based PPP and PPP-RTK as well as network-RTK services, yielding higher positioning accuracies. These services are not included in Table 26.7, as the table is restricted to DGPS based on code data only. The positioning accuracies that are claimed by the DGNSS code-based service providers are all in the same range: about 1–3 m, although some (commercial) providers claim submeter positioning accuracy.

Next to the correction message, many governmental/commercial DGNSS systems provide a form of *integrity* message to alert the users of a potential fault in either the corrections or the GNSS signals themselves. More details on the integrity monitoring of GNSS signals can be found in Chap. 24.

26.2.3 Data Communication: RTCM Message

The data communication link is essential for the (real-time) transmission of the DGNSS corrections to users. Most DGNSS service providers use a *standardized* format that is defined and published by the Radio Technical Commission for Maritime Services Special Committee 104 (RTCM SC-104) (Annex). Initially, RTCM corrections were defined only for pseudorange differential GPS (version 2.0) as this is all that is required to provide a marine DGPS service. Support for carrier-phase GPS measurements was added in version 2.1, GLONASS support was included in

version 2.2, and modernized data structures, increased bandwidth efficiency and support network RTK was implemented with version 3.0. Versions 2.3 and 3.2 are current though the standards continue to evolve to meet user needs. Both versions 2.3 and 3.2 define message formats for transmitting station information, pseudorange and range-rate corrections (Sect. 26.2.4) as well as carrier-phase raw data and carrier-phase corrections for GPS and GLONASS. Version 3.2 provides additional support, through multiple signals messages, for most of the emerging GNSSs as well as support for various formats of network RTK and PPP corrections [26.36].

While SBAS systems deploy (communication) satellites as link for the communication of the RTCM corrections, for GBAS systems there are several communication links in use:

- Radio communication (VHF/High frequency (HF); Frequency Modulation (FM) Radio Data Service (RDS))
- Mobile communication (Global System for Mobile Communication (GSM), General Packet Radio Service (GPRS), Universal Mobile Telecommunication System (UMTS))
- Internet communication (NTRIP).

Note that GPRS makes use of a combination of Internet and mobile telephone (GSM), whereas UMTS is the successor of GPRS. Of specific interest is the *networked transport of RTCM via Internet protocol* (NTRIP) for the streaming of DGNSS data which was developed by the German Federal Agency for Cartography and Geodesy (BKG) in 2004 [26.37]. The standard has been adopted by most receiver manufacturers allowing RTCM corrections to be broadcast over the Internet.

26.2.4 Latency of DGNSS Corrections

Users of DGNSS would like to have their corrections corresponding to the same time as the time at which they collect their measurements, in order to compute a differential solution with the best possible accuracy. In practice, however, there may be some delay or *latency* before the corrections arrive at the user. In addition, the reference station (or the network) may only transmit corrections at a certain time interval (i.e., the *update rate*), which may not correspond with the data sampling rate of the user.

To deal with this latency problem and to provide corrections that can be applied during the time before a new set of corrections arrives, the RTCM protocol uses a first-order polynomial to *predict* the DGNSS cor-

rections, based on the rate of these DGNSS corrections, the so-called *range-rate* corrections

$$\text{PRC}(t) = \text{PRC}(t_0) + (t - t_0)\text{RRC}(t_0). \quad (26.10)$$

Here t_0 denotes the time instant the pseudorange corrections (PRC) are determined at the reference station and t the time instant corresponding to the measurements of the user, such that their difference $t - t_0$ is the latency. The range-rate corrections are denoted by range-rate correction (RRC) and these are also part of the RTCM correction message, next to the pseudorange corrections. It will be clear that the acceptable latency depends on the application at hand and the DGNSS position accuracy that is required.

Before SA was turned off, the DGPS pseudorange corrections varied much more in time than after it was removed. With SA on, the RRCs were useful in

reducing the update rate of the DGPS corrections. However, with SA turned off, the temporal variations in the pseudorange corrections are predominantly governed by the temporal variations in the atmospheric and orbit errors. Hence, as the RRCs vary very slowly and are almost zero, some authors propose not to use it anymore as it may deteriorate the DGPS position accuracy [26.38].

Even if the update rate of the DGNSS corrections is high, there is always some latency due to the communication link that is used to transmit them to the users. For example, a DGPS RTCM data stream as broadcast by the EUREF station in Brussels to a user in Frankfurt via the NTRIP caster, resulted in latencies of up to one second for transmission through the Internet. Using GPRS, that is, a combination of mobile phone and the Internet, the latencies were at most two seconds [26.39].

26.3 Real-Time Kinematic Positioning

This section discusses the aspects of carrier-phase-dominated DGNSS positioning and in particular the RTK positioning technique that relies on resolving the carrier-phase ambiguities to estimate the receiver position with a high precision. Before RTK is discussed, Sect. 26.3.1 reviews the double-differenced observation equations underlying the carrier-phase and code-based positioning model. As the receiver-satellite geometry has a drastic impact on the precision of the estimated carrier-phase ambiguities and position coordinates, Sect. 26.3.2 is devoted to this and relates it to the development of carrier-phase-based positioning techniques from a historical point of view, starting from conventional static GNSS to RTK positioning. The last three subsections are all devoted to RTK. Section 26.3.3 discusses intricacies that occur when GLONASS data are used for RTK, whereas Sect. 26.3.4 addresses aspects that are relevant to multi-GNSS RTK positioning, including intersystem biases and inter-satellite-type biases. Finally, Sect. 26.3.5 presents examples of the performance of RTK ambiguity resolution and positioning.

26.3.1 Double-Differenced Positioning Model

When carrier-phase observations are included in the positioning model, the system of between-receiver differenced observation equations (26.1) cannot be used directly. As with DGNSS based on the code data, it is first assumed that the satellite positions are known ($\Delta \mathbf{r}^s = 0$). Also the position of the receiver position is

usually assumed to be known ($\Delta \mathbf{r}_1 = 0$), although this is strictly not required.

Having done so, the system can still not be solved in a unique way, as it is *rank-deficient* (Chap. 22), because the columns of the design matrix between the differential receiver phase biases and differential ambiguities are linear dependent. To overcome this rank deficiency, one of the satellites has to be selected as *pivot satellite*. Alternatively, the rank deficiency between receiver phase biases and ambiguities can be eliminated by differencing the between-receiver differenced observations with respect to those of this pivot satellite. As discussed in Chap. 21, this then leads to the *double-differenced* (linearized) observation equations for pseudorange and carrier-phase

$$\begin{aligned} \Delta p_{1r,j}^{1s} &= -\mathbf{e}_r^{1s\top} \Delta \mathbf{r}_{1r} + T_{1r}^{1s} + \mu_j^S I_{1r}^{1s} \\ &\quad + c \Delta d_{1r,j}^{1s} + e_{1r,j}^{1s}, \\ \Delta \varphi_{1r,j}^{1s} &= -\mathbf{e}_r^{1s\top} \Delta \mathbf{r}_{1r} + T_{1r}^{1s} - \mu_j^S I_{1r}^{1s} \\ &\quad + c \Delta \delta_{1r,j}^{1s} + \lambda_j^S N_{1r,j}^{1s} + \varepsilon_{1r,j}^{1s}. \end{aligned} \quad (26.11)$$

The double-differenced observables and parameters are denoted by $(\cdot)^{1s} = (\cdot)^s - (\cdot)^1$, with satellite 1 selected as pivot. For sufficiently short baselines, it is usually allowed to ignore the double-differenced tropospheric and ionospheric delays (Sect. 26.1.2). If they cannot be neglected, a common procedure for the tropospheric delays is to correct the observations using an a-priori troposphere model and map the residual delays to local zenith, and this zenith troposphere delay (ZTD)

is estimated as an unknown parameter. If the double-differenced ionospheric delays cannot be ignored as well, they are estimated as unknown parameters, together with the other parameters (Chap. 21).

26.3.2 Carrier-Phase-Based Positioning Methods

The carrier-phase observations have millimeter precision. However, high (millimeter to centimeter) precision for the position parameters can unfortunately not be attained directly, which is due to the presence of the unknown carrier-phase ambiguities, even when the differential atmospheric parameters can be ignored. In order to solve both position and ambiguities with high precision, one needs a *long* observation time (can be more than half an hour), which is due to the *receiver-satellite geometry* that changes only slowly, as GNSS satellites are in a very high altitude orbit with respect to an Earth-bound receiver.

Using a short-time span, the position and ambiguity precision is predominantly governed by the precision of the *pseudorange* data, which is at the level of a few decimeters for geodetic receivers. In the limiting case, with a *single epoch* of data, the position solution is fully determined by the pseudorange data, as the phase data are all needed to solve the ambiguity parameters. In the presence of multiple epochs of data, the phase data start to contribute to the position/ambiguity solution. However within a short-time span this contribution is only marginal, which can be understood as follows, following the reasoning provided in [26.25].

Conventional Static Positioning

Consider the following system of linearized observation equations for double-differenced *phase data only* of m satellites collected by a stationary receiver during k observation epochs

$$E \begin{bmatrix} y(t_1) \\ y(t_2) \\ \vdots \\ y(t_k) \end{bmatrix} = \begin{bmatrix} \mathbf{A}(t_1) & \mathbf{I} \\ \mathbf{A}(t_2) & \mathbf{I} \\ \vdots & \vdots \\ \mathbf{A}(t_k) & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \nabla \end{bmatrix}. \quad (26.12)$$

In the above system, $y(t_i)$ denotes a $f(m-1)$ vector that contains the $m-1$ double-differenced phase observations for each of the f frequencies at epoch i , \mathbf{x} denotes the vector of relative coordinate components and ∇ the vector of $f(m-1)$ carrier-phase ambiguities. The receiver-satellite geometry, as contained in the line-of-sight vectors at epoch i , is captured by $f(m-1) \times 3$ matrix $\mathbf{A}(t_i)$. Thus, for one frequency

$$\mathbf{A}(t_i) = [-\mathbf{e}_r^{12}(t_i), \dots, -\mathbf{e}_r^{1m}(t_i)]^\top.$$

Matrix \mathbf{I} denotes the identity matrix of dimension $f(m-1)$.

Note that the carrier-phase ambiguities are parameterized as multiplied by their wavelengths (so they are expressed in meters rather than cycles). These ambiguities do not have a time index, as they are constant in time, as long as no *cycle slips* occur. A cycle slip causes a jump of the ambiguity, but this jump (which equals an integer multiple of wavelengths) can be detected by application of the standard hypothesis testing theory ([26.40, 41] and Chap. 24). Having detected a cycle slip, the corresponding ambiguity parameter is adapted for it, such that the time-constancy of the ambiguity is not violated.

If the time that is spanned by the epochs 1 to k is short (e.g., a few minutes), the receiver-satellite geometry will only change slowly, as the line-of-sight vectors will not be very different from one epoch to another. In this case, $\mathbf{A}(t_1) \approx \mathbf{A}(t_2) \approx \dots \approx \mathbf{A}(t_k)$. In the limiting case, if the receiver-satellite geometry is assumed to be *stationary*, that is, $\mathbf{A}(t_1) = \mathbf{A}(t_2) = \dots = \mathbf{A}(t_k)$, the system of observation equations (26.12) is rank deficient, as the columns of the design matrix between position and ambiguities are linear dependent. Although, in practice, the receiver-satellite geometry is not stationary (except for geostationary satellites from e.g., BDS), the slow change of it has as consequence that both position and ambiguities are only poorly estimable. Hence, in order to be able to estimate \mathbf{x} and ∇ with sufficient precision, one will have to make sure that the receiver-satellite geometry has changed significantly, such that $\mathbf{A}(t_1) \neq \mathbf{A}(t_k)$. Using the conventional static GPS surveying technique [26.42, 43] it takes typically 20–30 min for short baselines (ignoring differential ionospheric delays), but it could take several (1–3) hours for long baselines (parameterizing the differential ionospheric delays as unknown parameters) [26.25].

Semikinematic Positioning

As the conventional static positioning technique is not very attractive in terms of productivity due to its long observation times, several positioning techniques have been proposed in the past to reduce the observation time that is needed to solve the position (and ambiguities) with sufficient precision [26.44, 45].

Three variants of these *semikinematic* positioning techniques are discussed here (Fig. 26.6):

1. With revisiting of stations [26.46]
2. Starting from a known baseline
3. With antenna swap.

All the three methods are based on the assumption that the rover receiver is allowed to collect data for a short

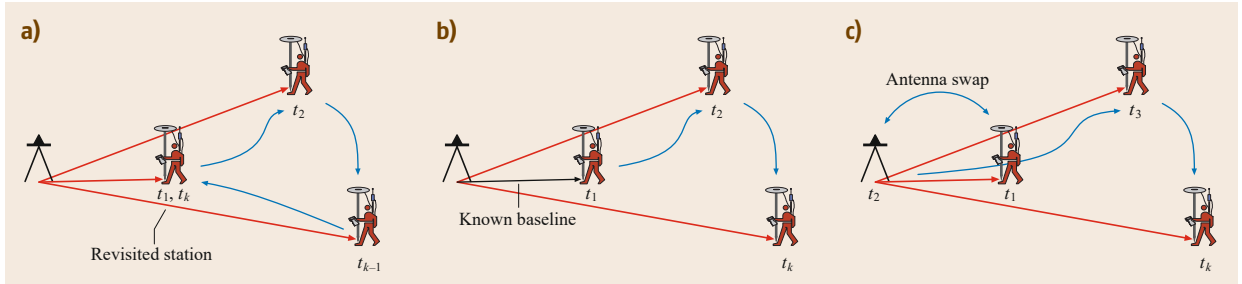


Fig. 26.6a–c Semikinematic relative positioning strategies: **(a)** with revisiting of stations; **(b)** starting from a known baseline; **(c)** with antenna swap

time (e.g., a few minutes) at one point and then visits another point to collect data for only a short time as well. In this way, the productivity can be increased compared to static positioning. Instead of model (26.12), these semikinematic methods are based on the following model

$$E \begin{pmatrix} \mathbf{y}(t_1) \\ \vdots \\ \mathbf{y}(t_k) \end{pmatrix} = \begin{bmatrix} \mathbf{A}(t_1) & & \mathbf{I} \\ & \ddots & \vdots \\ & & \mathbf{A}(t_k) & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}(t_1) \\ \vdots \\ \mathbf{x}(t_k) \\ \nabla \end{bmatrix}. \quad (26.13)$$

Here $\mathbf{x}(t_i)$ denotes the relative position vector at epoch i . Thus, provided that the area that is covered is of small scale, the rover receiver collects data during a short time at each point, continuously tracking the satellites when moving between the points.

With Revisiting of Stations. With conventional static positioning, it is not so much the number of epochs

but the *change of geometry* that contributes to the determination of the phase ambiguities. Hence, one can suffice with two periods of data collection at the same point in static mode, where these two periods are separated by a sufficient long time interval (e.g., 30 min if the distance between reference and rover is short). During this time interval, the rover receiver can visit other points. In Table 26.8, the model corresponding to this *revisiting stations* semikinematic positioning technique is presented, where it is assumed that the first point is observed for a short time at both the start and end of the session, such that the geometry has changed considerably, that is, $\mathbf{A}(t_1) \neq \mathbf{A}(t_k)$.

Starting from a Known Baseline. Another variant of semikinematic positioning requires two points with *precisely known coordinates*, at which the reference and rover receivers are placed at the first epoch. As the relative position is known, the ambiguities can be quickly determined with very high precision. After this initialization, the rover receiver is moved to a next point for quick position determination, thereby continuously

Table 26.8 System of observation equations corresponding to the semikinematic relative positioning strategies

Variant	Model
With revisiting of stations	$E \begin{pmatrix} \mathbf{y}(t_1) \\ \vdots \\ \mathbf{y}(t_{k-1}) \\ \mathbf{y}(t_k) \end{pmatrix} = \begin{bmatrix} \mathbf{A}(t_1) & & \mathbf{I} \\ & \ddots & \vdots \\ & & \mathbf{A}(t_{k-1}) & \mathbf{I} \\ \mathbf{A}(t_k) & & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}(t_1) \\ \vdots \\ \mathbf{x}(t_{k-1}) \\ \nabla \end{bmatrix}$
Starting from a known baseline	$E \begin{pmatrix} \mathbf{y}(t_1) - \mathbf{A}(t_1)\mathbf{x}(t_1) \\ \mathbf{y}(t_2) \\ \vdots \\ \mathbf{y}(t_k) \end{pmatrix} = \begin{bmatrix} \mathbf{0} & & \mathbf{I} \\ \mathbf{A}(t_2) & & \mathbf{I} \\ & \ddots & \vdots \\ & & \mathbf{A}(t_k) & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}(t_2) \\ \vdots \\ \mathbf{x}(t_k) \\ \nabla \end{bmatrix}$
With antenna swap	$E \begin{pmatrix} \mathbf{y}(t_1) \\ \mathbf{y}(t_2) \\ \mathbf{y}(t_3) \\ \vdots \\ \mathbf{y}(t_k) \end{pmatrix} = \begin{bmatrix} \mathbf{A}(t_1) & & \mathbf{I} \\ -\mathbf{A}(t_2) & & \mathbf{I} \\ & \mathbf{A}(t_3) & \mathbf{I} \\ & & \ddots & \vdots \\ & & & \mathbf{A}(t_k) & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}(t_1) \\ \mathbf{x}(t_3) \\ \vdots \\ \mathbf{x}(t_k) \\ \nabla \end{bmatrix}$

tracking the signals. The model corresponding to this technique is given in Table 26.8, where the relative position at the first epoch is known, such that the observations at the first epoch are corrected for this. Note that the design matrix is now of full rank, even in the case $\mathbf{A}(t_1) = \dots = \mathbf{A}(t_k)$.

With Antenna Swap. In this variant, the idea is to determine the ambiguities with high precision by moving the antenna of the reference receiver to the initial location of the rover receiver, while, at the same time, the antenna of the rover receiver is moved to the location of the reference receiver. This implies that after the antennas have been swapped, the carrier-phase ambiguities are not changed (as the satellites are continuously tracked), but the relative position after the swap is of *opposite sign* as before the swap. Making use of this property, that is, $\mathbf{x}(t_2) = -\mathbf{x}(t_1)$, results in the *antenna swap* model as given in Table 26.8. Also this model is of full rank, even in case $\mathbf{A}(t_1) = \dots = \mathbf{A}(t_k)$.

(Real-Time) Kinematic Positioning

Instead of waiting until the (float) ambiguities are precise enough such that the position can be estimated with high precision, one can make use of the *integerness* of the double-differenced carrier-phase ambiguities. Once the ambiguities can be resolved to their integer values, they can be removed from the system of observation equations, such that the position can be solved with a very high precision.

The procedure to solve for the precise position is as follows (Chap. 23). One starts to estimate the ambiguities using model (26.13) using standard least-squares or Kalman filtering. This solution is referred to as *float solution*. Although for short-time spans the float ambiguities have a poor precision and high correlation, it is possible to *decorrelate* them resulting in ambiguities that have a better precision and less correlation. By means of a special *search technique* it is then possible to find the integer values of the double-differenced ambiguities. Both decorrelation and integer search are efficiently implemented in the least-squares ambiguity decorrelation adjustment (LAMBDA) method, which is the standard for *integer ambiguity resolution* (Chap. 23). Having resolved the integer ambiguities, a second standard least-squares estimation is carried out keeping these ambiguities fixed in the model

$$E \left(\begin{bmatrix} \mathbf{y}(t_1) - \nabla \\ \vdots \\ \mathbf{y}(t_k) - \nabla \end{bmatrix} \right) = \begin{bmatrix} \mathbf{A}(t_1) & & \\ & \ddots & \\ & & \mathbf{A}(t_k) \end{bmatrix} \begin{bmatrix} \mathbf{x}(t_1) \\ \vdots \\ \mathbf{x}(t_k) \end{bmatrix}. \quad (26.14)$$

The solution if this model is referred to as *fixed solution*. The fixed solution has high precision as the carrier-phase data with the integer ambiguities subtracted now act as *very precise pseudorange* data.

The success of integer ambiguity resolution depends on the *strength* of the underlying positioning model. *Instantaneous ambiguity resolution*, where the integer ambiguities can be resolved already after a single epoch of data, is only feasible for very strong models, such as the short-baseline GPS model based on dual-frequency phase and code data [26.47]. Weaker models, such as the short-baseline single-frequency model or the long-baseline model parameterizing ionospheric delays [26.48], require more time before the ambiguities can be reliably fixed. These weaker models can however be strengthened by incorporating dynamic models (in a Kalman filter) on some of the parameters (i. e., position, ZTD, ionospheric delays, hardware biases), or, alternatively, by combining data of multiple GNSS constellations (Sect. 26.3.4).

Because of its high productivity provided that the distance between reference and rover is short, RTK is commonly used in cadastral and engineering surveying.

26.3.3 GLONASS RTK Positioning

The double-differenced observation equations (26.11) form the basis of the model underlying RTK positioning. It is remarked that a formulation in terms of the *undifferenced* observation equations may however be used as well (see Chap. 21 as well as [26.49, 50]). In this section, we will address the issues that arise when the double-differenced observation equations are used for RTK based on GLONASS observations.

For GLONASS signals that are based on the FDMA technology (Chap. 8), we have to deal with interchannel code and phase biases, denoted by $\Delta d_{1r,j}^s$ and $\Delta \delta_{1r,j}^s$ in (26.1), which are not automatically eliminated in the double-differenced positioning model (26.11). A usual assumption for the GLONASS interchannel phase biases is that they can be modeled as a *linear* function of the frequency or channel [26.51], that is, $\Delta \delta_{1r,j}^s = \kappa^s \Delta \delta_{1r,j}$, with κ^s denoting the *integer channel number*. Based on this, for a short baseline (neglecting the differential atmospheric biases), the double-differenced observation equations for GLONASS read

$$\begin{aligned} \Delta p_{1r,j}^{1s} &= -\mathbf{e}_r^{1s\top} \Delta \mathbf{r}_{1r} + c \Delta d_{1r,j}^{1s} + e_{1r,j}^{1s}, \\ \Delta \phi_{1r,j}^{1s} &= -\mathbf{e}_r^{1s\top} \Delta \mathbf{r}_{1r} + \kappa^{1s} c \Delta \delta_{1r,j} \\ &\quad + \lambda_j^s \left(N_{1r,j}^s - \frac{\lambda_j^1}{\lambda_j^s} N_{1r,j}^1 \right) + \varepsilon_{1r,j}^{1s}. \end{aligned} \quad (26.15)$$

In contrast to the phase interchannel biases, the inter-channel biases for code can however not be modeled as a linear function [26.52]. Besides the presence of the differential interchannel biases, there is another difference with the phase observation equations for CDMA signals: the carrier-phase ambiguity is due to the satellite or channel-dependent wavelengths *not directly estimable as a double-differenced ambiguity*, which would imply that we cannot make use of its integer property, which is the key to high-precision RTK positioning.

The presence of phase interchannel biases poses a first problem for the estimation of the ambiguities, as they cannot be estimated separately from each other (as their columns in the design matrix of the model would show a rank deficiency). However, in case the RTK baseline is formed by pairs of receivers that are of the *same manufacturer*, it turns out that they can be ignored, that is, $\Delta\delta_{1r,j}^{1s} = 0$ and $\Delta\delta_{1r,j} = 0$, removing them from the differential GLONASS observation equations [26.53]. In the case that the baseline is observed by receivers of *different manufacturers* (i.e., mixed receiver pairs), it is possible to calibrate these inter-channel biases, based on their stability [26.54]. A-priori correction values for $\Delta\delta_{1r,j}$ are provided in [26.54, 55], from which follows that they can range up to 5 cm for adjacent channels (i.e., for $\kappa^{1s} = 0$) for a certain mixed receiver pair. The size of the interchannel biases for code may range up to 5 m [26.56].

Unfortunately, ignoring or correcting the inter-channel biases does not automatically mean that the GLONASS ambiguities can be estimated as double-differences and thus as integers. In the literature, some authors [26.53, 57] try to overcome this by rewriting the ambiguity term as its equivalent

$$\lambda_j^s \left(N_{1r,j}^s - \frac{\lambda_j^1}{\lambda_j^s} N_{1r,j}^1 \right) = \lambda_j^s N_{1r,j}^{1s} + \lambda_j^{1s} N_{1r,j}^1. \quad (26.16)$$

Thus, the GLONASS ambiguity combination is rewritten as a double-differenced ambiguity $N_{1r,j}^{1s}$, plus the between-receiver differenced ambiguity that corresponds to the pivot satellite $N_{1r,j}^1$. Unfortunately the two ambiguity terms cannot be estimated as separate parameters. Takac [26.53] proposes to get an approximation of $N_{1r,j}^1$ by estimating it from the difference of the between-receiver differenced code and phase observations

$$\begin{aligned} \Delta p_{1r,j}^1 - \Delta \phi_{1r,j}^1 &= c [d_{1r,j}^r - \delta_{1r,j}^r] - \lambda_j^1 N_{1r,j}^1 \\ &+ e_{1r,j}^1 - \varepsilon_{1r,j}^1. \end{aligned} \quad (26.17)$$

The additional assumption to approximate $N_{1r,j}^1$ is that differential code-phase receiver bias $d_{1r,j}^r - \delta_{1r,j}^r$ is zero

for identical receiver pairs, while it can be calibrated for mixed receiver pairs. However, since the precision of the approximate $N_{1r,j}^1$ is not very high, as it is driven by the noisy code data, it may result in unreliable resolution of the double-differenced ambiguities $N_{1r,j}^{1s}$ [26.58].

An alternative formulation to deal with the GLONASS pivot ambiguity is presented by [26.59]. Here the authors reparameterize the GLONASS ambiguity terms in (26.16) as

$$\lambda_j^s N_{1r,j}^{1s} + \lambda_j^{1s} N_{1r,j}^1 = \lambda_j^s \tilde{N}_{1r,j}^{1s} + \lambda_j^{1s} \tilde{N}_{1r,j}^1. \quad (26.18)$$

Here the double-differenced and pivot satellite ambiguities are reparameterized as

$$\begin{aligned} \tilde{N}_{1r,j}^{1s} &= N_{1r,j}^{1s} - \frac{\kappa^{1s}}{\kappa^{12}} N_{1r,j}^{12}, \quad s \geq 3 \\ \tilde{N}_{1r,j}^1 &= N_{1r,j}^1 + \frac{\lambda_j^2}{\lambda_j^{12}} N_{1r,j}^{12}. \end{aligned} \quad (26.19)$$

Use is made of the property that

$$\lambda_j^s \frac{\kappa^{1s}}{\kappa^{12}} = \lambda_j^2 \frac{\lambda_j^{1s}}{\lambda_j^{12}}.$$

Note that the reparameterized double-differenced ambiguity $\tilde{N}_{1r,j}^{1s}$ is biased by a multiple of the integer ambiguity between satellite 2 and the pivot satellite, that is, $N_{1r,j}^{12}$. Because of this, satellite 2 can be considered as a *second pivot satellite*. Because the reparameterized double-differenced ambiguity $\tilde{N}_{1r,j}^{1s}$ is not an unknown parameter for this second pivot satellite (it only shows up from the third satellite onward), it is now possible to separate both reparameterized ambiguities, which means that *both* double-differenced ambiguities $\tilde{N}_{1r,j}^{1s}$ and the between-receiver differenced ambiguity of the first pivot satellite $\tilde{N}_{1r,j}^1$ can be estimated.

There seems however to be one problem: the double-differenced ambiguities $\tilde{N}_{1r,j}^{1s}$ are generally not integer, as they are biased by $(\kappa^{1s}/\kappa^{12})N_{1r,j}^{12}$. Although the GLONASS channel numbers κ^s are integers (and thus κ^{1s} as well), in general the fraction κ^{1s}/κ^{12} destroys the integerness of $\tilde{N}_{1r,j}^{1s}$. However, there is an exception: in case the channel numbers of the two pivot satellites differ by one (i.e., $|\kappa^{12}| = 1$), the fraction κ^{1s}/κ^{12} will actually be an integer and, consequently $\tilde{N}_{1r,j}^{1s}$ will be an integer as well. Thus, this implies that the two GLONASS pivot satellites cannot be arbitrarily chosen, but this should be two satellites with *adjacent channel numbers*. Only in that case integer resolution of GLONASS ambiguities becomes possible. However, it is stressed that this approach of GLONASS integer ambiguity resolution fully relies on the underlying assumption that the differential interchannel biases may

be ignored (i. e., $\Delta d_{1r,j}^{1s} = 0$ and $\Delta \delta_{1r,j} = 0$), which only holds for identical receiver pairs. For mixed receivers still external calibrations are needed, despite the claims made by [26.59].

26.3.4 Multi-GNSS RTK Positioning

The presence of more than one GNSS constellation benefits RTK positioning as code and phase data of multiple systems can be integrated in the positioning model. This was already demonstrated in Chap. 21 for the model underlying SPP, but it also applies to RTK positioning.

If we restrict ourselves to CDMA-based constellations, the interchannel biases are absent and for an arbitrary constellation S the between-receiver differenced code and phase observation equations read as in (26.1). Selecting a *pivot satellite* for each constellation then yields double-differenced observation equations, similar to (26.11), but then for each constellation. Common parameters between the constellations are the receiver position coordinates, plus ZTDs if these are parameterized; all other estimable parameters are constellation specific. It is possible to have a multi-GNSS model that is stronger than sketched as above, by making use of frequencies that are *identical* between different systems (e.g., GPS L1 and Galileo E1).

Differential Intersystem Biases (DISBs)

For frequencies that are identical between constellations, it is possible to difference the data of the second constellation using the data of the pivot satellite of the *first* constellation. Consider two receivers tracking data of two constellations, denoted by A and B (Fig. 26.7), then the double-differenced observation equations read (for a short baseline), for observations of constellation A

$$\begin{aligned} E(\Delta p_{1r,j}^{1A s_A}) &= -\mathbf{e}_r^{1A s_A^\top} \Delta \mathbf{r}_{1r}, \\ E(\Delta \varphi_{1r,j}^{1A s_A}) &= -\mathbf{e}_r^{1A s_A^\top} \Delta \mathbf{r}_{1r} + \lambda_j^A N_{1r,j}^{1A s_A}. \end{aligned} \quad (26.20)$$

Here $s_A = 2_A, \dots, m_A$, with m_A the number of satellites that are tracked of constellation A. The pivot satellite of A is denoted by 1_A . We may now difference the observations of constellation B that are tracked on the same frequency as observations of constellation A relative to the pivot satellite of A. This yields, assuming $\lambda_j^B = \lambda_j^A$

$$\begin{aligned} E(\Delta p_{1r,j}^{1A s_B}) &= -\mathbf{e}_r^{1A s_B^\top} \Delta \mathbf{r}_{1r} + c d_{1r,j}^{AB}, \\ E(\Delta \varphi_{1r,j}^{1A s_B}) &= -\mathbf{e}_r^{1A s_B^\top} \Delta \mathbf{r}_{1r} + c \delta_{1r,j}^{AB} \\ &\quad + \lambda_j^A N_{1r,j}^{1A s_B}. \end{aligned} \quad (26.21)$$

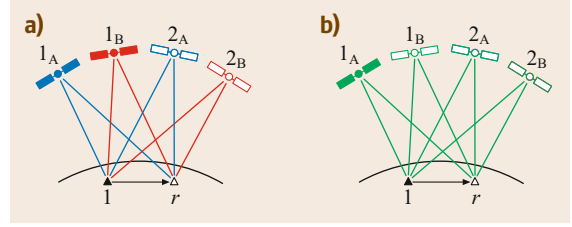


Fig. 26.7 (a) Two constellations, A and B, each defining its own pivot satellite, denoted by 1_A and 1_B ; (b) two constellations, A and B, having the pivot satellite in common, in this case 1_A , based on the assumption that the DISBs are known or zero

Here $s_B = 1_B, \dots, m_B$, with m_B the number of satellites that are tracked of constellation B. Note that for constellation B, we have one double difference more for phase and code, which corresponds to the first satellite of B, denoted by 1_B . At the same time, there are additional parameters to be estimated for constellation B, which are

$$d_{1r,j}^{AB} = d_{1r,j}^B - d_{1r,j}^A \quad \text{and} \quad \delta_{1r,j}^{AB} = \delta_{1r,j}^B - \delta_{1r,j}^A. \quad (26.22)$$

These are the *differential intersystem biases* (DISBs) [26.60] for code and phase. We emphasize that the above definitions of the DISB parameters apply to the short-baseline model in the absence of differential atmospheric biases. The estimability and interpretation of the DISBs change for the ionosphere-float model, in which the ionospheric delays are estimated and which is used for longer baselines [26.61].

The performance of the model parameterized into DISBs is exactly the same as model with constellation-specific pivot satellites. Only in case we have information on these DISBs and include this in the model, the performance will be better. In the case we *a priori know* the DISBs, the observations can be corrected for them

$$\begin{aligned} E(\Delta p_{1r,j}^{1A s_B} - c d_{1r,j}^{AB}) &= -\mathbf{e}_r^{1A s_B^\top} \Delta \mathbf{r}_{1r}, \\ E(\Delta \varphi_{1r,j}^{1A s_B} - c \delta_{1r,j}^{AB}) &= -\mathbf{e}_r^{1A s_B^\top} \Delta \mathbf{r}_{1r} + \lambda_j^A N_{1r,j}^{1A s_B}. \end{aligned} \quad (26.23)$$

for $s_B = 1_B, \dots, m_B$. Now, in the absence of DISB parameters, the observations of constellation B can be processed identically as those of constellation A, as they have a common pivot satellite and identical type of parameters. In other words, both constellations can be processed *as if* their signals correspond to one constellation (Fig. 26.7). This is referred to as *tightly combined* processing in [26.62]. Whereas in a combined model consisting of observation equations (26.11) for each constellation, there are $f(m_A + m_B - 2)$ estimable integer ambiguities (assuming f frequencies for both constella-

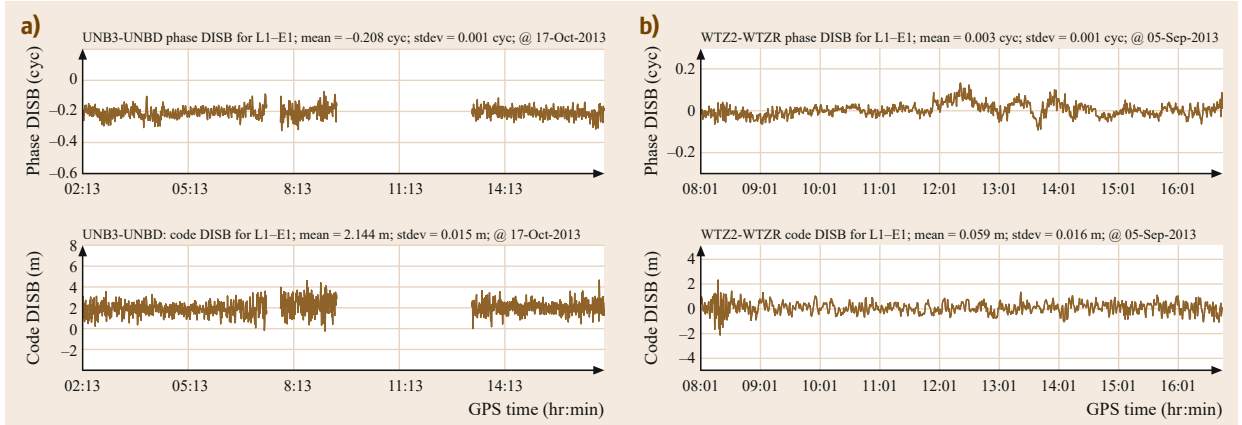


Fig. 26.8a,b Example of L1 (GPS) – E1 (Galileo) DISBs for phase (*top graphs*) and code (*bottom graphs*), estimated for the 20 m baseline between UNB3 (Trimble NetR9) and UNBD (Javad TRE-G2T) (**a**) and the zero baseline between WTZ2 (Leica GR25) and WTZR (Leica GRX1200+GNSS) (**b**). Gaps in the time series mean that Galileo satellites were not tracked during that period

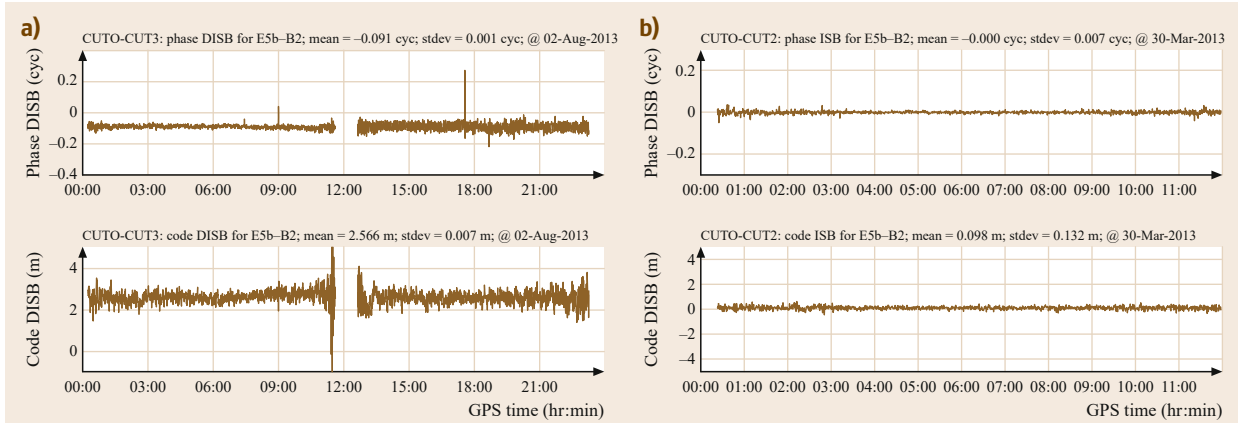


Fig. 26.9a,b Example of E5b (Galileo) – B2 (BDS) DISBs for phase (*top graphs*) and code (*bottom graphs*), estimated for the zero baseline between CUTO (Trimble NetR9) and CUT3 (Javad TRE-G3TH) (**a**) and the zero baseline between CUTO (Trimble NetR9) and CUT2 (Trimble NetR9) (**b**). Gaps in the time series mean that Galileo satellites were not tracked during that period

tions), this is increased with f integer ambiguities for the *tightly* combined or DISB-known model.

Concerning the size and variability of the DISBs, it was demonstrated that for pairs of *identical receivers*, that is, receivers of the same manufacturer, the DISBs are close to zero, that is, $\delta_{1r,j}^{AB} = 0$ and $d_{1r,j}^{AB} = 0$, whilst for *mixed receiver* combinations they are nonzero, but very stable in time [26.60, 63, 64].

As example, Fig. 26.8 depicts DISBs for phase and code that are estimated for the GPS L1 and Galileo E1 frequencies (both 1575.42 MHz). The figure shows that for the baseline at the University of New Brunswick, Canada, that consists of mixed receivers the mean of the estimated phase DISBs is about -0.2 cyc, while the estimated code DISB has a mean of about 2 m. For an-

other baseline in Wetzell, Germany, which consists of receivers that are both of Leica, the DISBs for both phase and code are estimated with a mean that is close to zero.

Figure 26.9 shows the estimated DISBs for another frequency, that is, Galileo's E5b and BDS's B2 frequency (both 1207.14 MHz). The left graphs show that for a zero baseline at the campus of Curtin University (Australia) consisting of mixed receivers the mean of the estimated phase DISBs is about -0.1 cyc, while the estimated code DISB has a mean of about 2.6 m. On the other hand, the right graphs depict that the phase and code DISBs estimated for another zero baseline at Curtin University, this time consisting of two identical Trimble receivers, are close to zero.

Differential Inter-Satellite-Type Biases (DISTBs)

In case BDS data are used for RTK positioning, either standalone or combined with other GNSSs, another type of bias has shown up in case the baseline consists of mixed receivers. This is the so-called (differential) inter-satellite-type bias (ISTB), that is present between the signals at the *same frequency* of the BDS geostationary (GEO) satellites on one hand and the other BDS (IGSO, MEO) satellites on the other hand [26.65, 66].

Table 26.9 summarizes the phase DISTBs for two receiver manufacturers, relative to a Trimble receiver. Like the DISBs, the DISTBs turn out to be very stable, such that they can be calibrated. It follows from [26.65] that the GEO satellites have phase DISTBs of exactly *half cycles* with respect to IGSO/MEO satellites in the case of mixed receivers. However, it depends on the combination of mixed receiver types, which frequency is actually affected and which not. As follows from Table 26.9, the B1 frequency is not affected for the Trimble–Septentrio combination, but is affected for the Trimble–Javad combination. For the B2 frequency, the effect is vice versa.

After the discovery in [26.65], the GNSS receiver manufacturers updated their receivers firmware in order to eliminate the DISTBs when using mixed receivers for BDS RTK [26.66]. However, users who employ receivers with old firmware should still be aware of the presence of DISTBs when processing BDS data.

26.3.5 RTK Positioning Examples

To provide insight into the actual performance of RTK ambiguity resolution and positioning, in this subsection examples are presented that are obtained in Western Australia. This is done based on data for both *GPS standalone* as well as *multi-GNSS RTK* (GPS+BDS).

Results are shown for both a short (1 km) baseline, observed between stations CUT0 and CUTT, both at Curtin University campus, as well as a long baseline that was measured between Curtin’s CUT0 and a receiver stationed in Muresk (MURK) at a distance of 80 km. *Short* and *long* here refer to whether the differential ionospheric biases can be neglected or not.

Table 26.9 Differential inter-satellite-type biases (DISTBs) (cyc) for phase between BDS geostationary Earth orbit (GEO) and inclined geosynchronous orbit (IGSO)/medium earth orbit (MEO) satellites with Trimble as pivot receiver

BDS frequency	Septentrio	Javad
B1	0	0.5
B2	0.5	0

Table 26.10 summarizes some details concerning the types of receivers that have collected the data, as well as the time of measurements, data cut-off angles, and sampling intervals. For the short baseline, data are processed in both *single-frequency* and *dual-frequency* modes, whereas for the long-baseline data are only processed in dual-frequency mode, as the long-baseline single-frequency model is too weak.

Once the float ambiguities are precise enough, integer ambiguity resolution is performed at each epoch by means of the LAMBDA method in combination with the fixed failure-rate ratio-test (FFRT; Chap. 23).

Short-Baseline RTK Results

Figures 26.10–26.14 show the time series of the position solution for rover CUTT, estimated in kinematic mode (for each epoch; no dynamic model on the position), in east-north-up. These components are relative to the ground-truth position for CUTT. Each figure shows two curves, one with the *ambiguities float* and the other with the *ambiguities fixed*. Next to the graphs, tables with empirical standard deviations are given. The float solution is obtained by means of a Kalman filter, with only a dynamic model on the ambiguities (i.e., they are time constant). Note from the figures that the ambiguity-fixed solution does not always start at the beginning of the time span, but it is only computed as soon as the ambiguities can be reliably fixed to integers (which is after acceptance by the FFRT). The tables with empirical position standard deviations therefore only present the float solution for the time after the ambiguities can be fixed, as to directly compare them to their fixed counterparts. This implies that one has to be careful when the float standard deviations are compared that correspond to different processing strategies.

From Figs. 26.10–26.13, it can be seen that in all cases the float solution needs time before it has converged, where the time of convergence depends on the strength of the model. If we assume a convergence criterion of 1 cm in both horizontal components and of 2 cm in the vertical component, Table 26.11 gives the time that is needed for each scenario in order for the float solution to converge.

Table 26.10 Characteristics of the short-baseline and long-baseline RTK examples

Short-baseline RTK	Long-baseline RTK
CUT0-CUTT: 1 km	CUT0-MURK: 80 km
2 Trimble NetR9 receivers	2 Trimble NetR9 receivers
Date: 20 April 2013	Date: 19 February 2014
Cutoff elevation: 10°	Cutoff elevation: 10°
Sampling interval: 30s	Sampling interval: 30 s

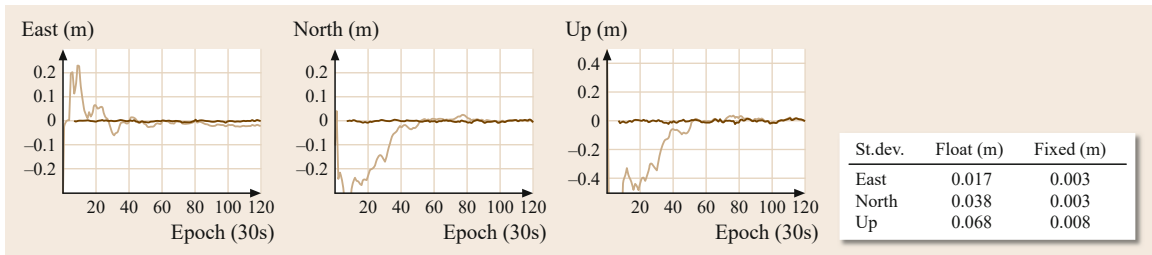


Fig. 26.10 GPS L1 RTK float (*light*) vs. fixed (*dark*) position errors, as well as standard deviations for 1 km baseline CUT0-CUTT

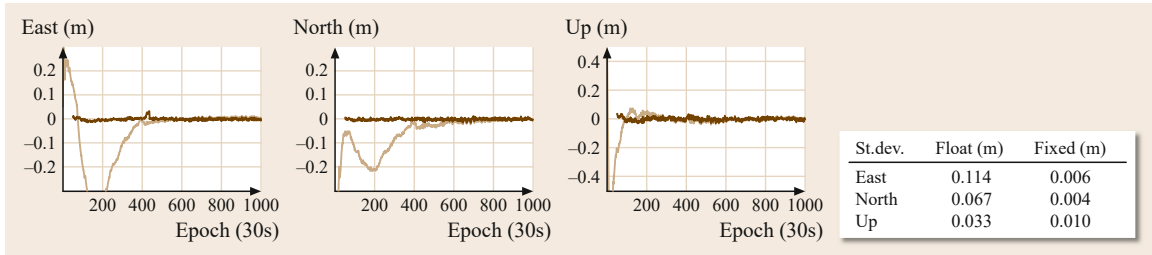


Fig. 26.11 BDS B1 RTK float (*light*) vs. fixed (*dark*) position errors, as well as standard deviations for 1 km baseline CUT0-CUTT

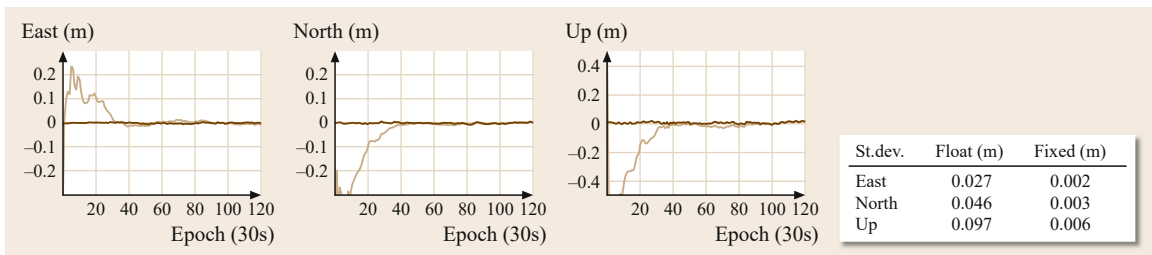


Fig. 26.12 GPS+BDS L1+B1 RTK float (*light*) vs. fixed (*dark*) position errors, as well as standard deviations for 1 km baseline CUT0-CUTT

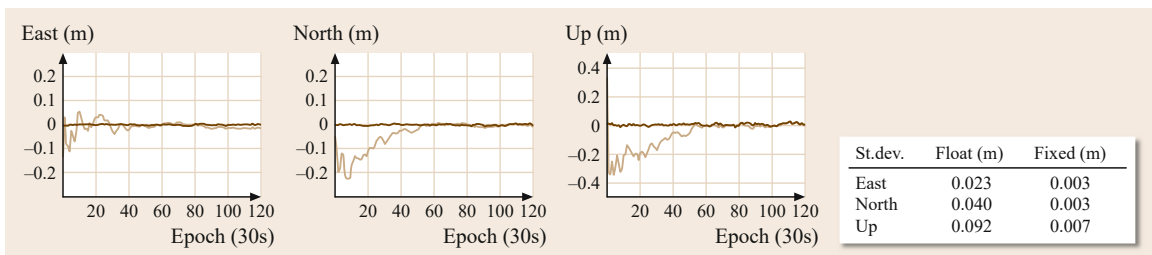


Fig. 26.13 GPS L1+L2 RTK float (*light*) vs. fixed (*dark*) position errors, as well as standard deviations for 1 km baseline CUT0-CUTT

For single-frequency GPS, this convergence time is 0.5 h, but for single-frequency BDS this is much longer: as many as 4 h are required. This very slow convergence is probably due to the geostationary satellites of BDS for which the geometry hardly changes over time. Moreover, due to this almost stationary geometry multipath on the signals of these geostationary satel-

lites gets barely averaged out in time [26.67]. With stronger models, the convergence time is reduced: for single-frequency GPS+BDS as well as dual-frequency GPS this is 20 min and 25 min, respectively (compared to 30 min for GPS L1). These relatively long convergence times can be avoided using *ambiguity resolution*.

Table 26.11 Performance of ambiguity resolution and positioning for 1 km baseline CUT0-CUTT (ionosphere-fixed model). The float position is converged if the horizontal position error < 1 cm and the vertical position error < 2 cm

Observables	Float position conv. after (min)	Fixed ambiguities/position after
GPS L1	30	4 min
BDS B1	240	25 min
GPS L1 & BDS B1	20	Instantaneous
GPS L1+L2	25	Instantaneous

Table 26.11 also presents the time that is needed to fix the ambiguities to integer. For single-frequency GPS this is 4 min and this is considerably less time than the 30 min that are needed for the float solution to converge. After these 4 min, the fixed position solution has immediately an accuracy at the centimeter level (Fig. 26.10). Also the time-to-fix-ambiguities for single-frequency BDS is much shorter than the float convergence time. For the stronger multi-GNSS (GPS L1 & BDS B1) and dual-frequency GPS models, the times-to-fix-ambiguities are extremely fast; based on only a single epoch of data the ambiguities can be fixed (*instantaneous* ambiguity resolution).

Alternatively, Fig. 26.14 presents the position errors for the dual-frequency GPS case, but now based on an *epoch-by-epoch* processing instead of Kalman filtering. In this case, the model is so strong such that it is not necessary to keep the ambiguities constant in time. As there are no dynamic models for any of the parameters, the float position solution is fully governed by the code data; this explains the noisy (decimeter to meter level) float position errors in Fig. 26.14. The fixed solution is however very precise (millimeter to centimeter level).

Then in this example the demonstrated performance of single-frequency multi-GNSS RTK even allows for an increase of the *cut-off elevation*, other than the customary cut-off angle that is typically set to 10° as done in this example. In [26.68, 69], it is demonstrated that a high cut-off angle set to 35° still results in high ambi-

Table 26.12 Performance of ambiguity resolution and kinematic positioning for 80 km baseline CUT0-MURK (ionosphere-float model). The float position is converged if the horizontal position error 2 cm and the vertical position error < 5 cm

Observables	Float position conv. after (min)	Fixed ambiguities/position after (min)
GPS L1+L2	185	35
BDS B1+B2	262	80
GPS L1+L2 & BDS B1+B2	165	20

guity resolution success rates. This makes RTK more robust for applications in environments where GNSS signals at low elevations are obstructed or experience multipath, such as for example in urban canyons or open-pit mines.

Long-Baseline RTK Results

In a similar way as for the short baseline, Figs. 26.15–26.17 show the float and fixed position errors but now for rover MURK corresponding to the long (80 km) baseline. Figure 26.15 gives the results for dual-frequency GPS, Fig. 26.16 for dual-frequency BDS and Fig. 26.17 for the combination of both constellations. For these three scenarios, Table 26.12 presents the time needed before the float solution is converged, as well as the time that is needed to reliably fix the integers and obtain a fixed solution.

The long-baseline model is weaker than the short-baseline model, because of the presence of differential ionospheric parameters, as well as a (residual) ZTD parameter. Because of this, the convergence criterion is slightly relaxed compared to the short-baseline model: 2 cm for the horizontal components and 5 cm for the vertical component. From Table 26.12, it follows that in case of GPS-only this criterion is met only after about 3 h. An even longer convergence time is required for BDS-only: almost 4.5 h. This longer time than GPS for BDS is, like in the short-baseline case, probably due to the stationary geometry of the geostationary satellites,

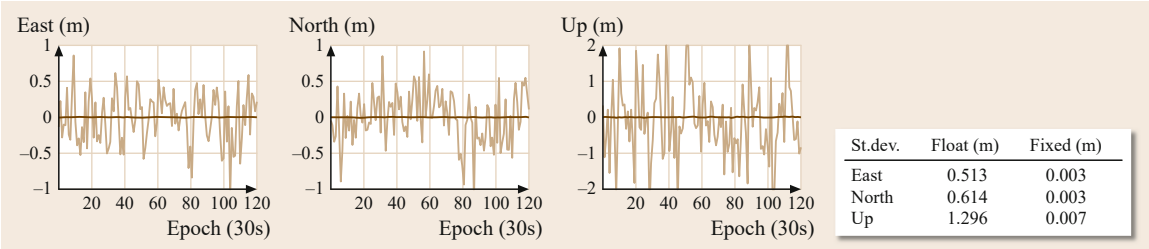


Fig. 26.14 GPS L1+L2 RTK float (*light*) vs. fixed (*dark*) position errors, as well as standard deviations for 1 km baseline CUT0-CUTT, based on *epoch-by-epoch* processing

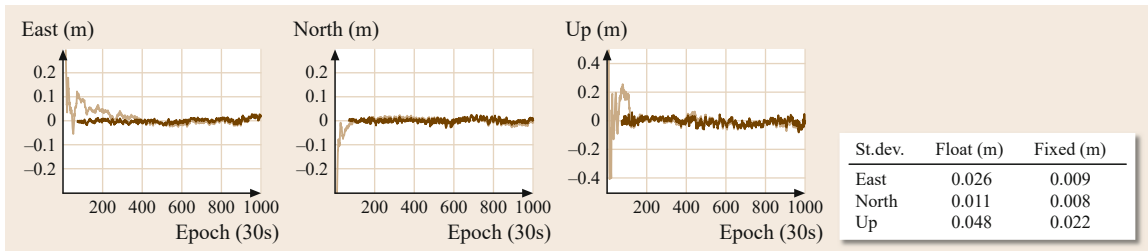


Fig. 26.15 GPS L1+L2 RTK float (*light*) vs. fixed (*dark*) position errors, as well as standard deviations for 80 km baseline CUT0-MURK

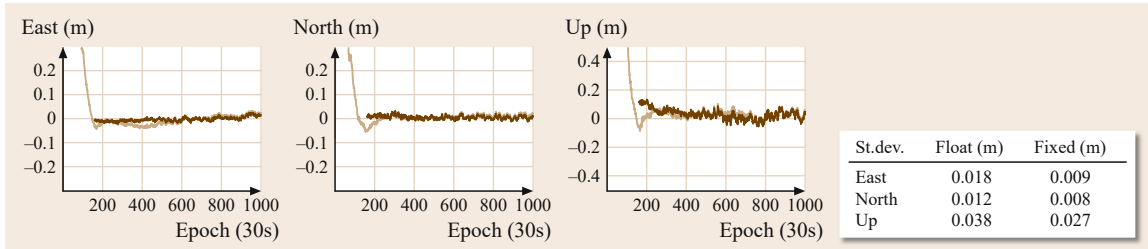


Fig. 26.16 BDS B1+B2 RTK float (*light*) vs. fixed (*dark*) position errors, as well as standard deviations for 80 km baseline CUT0-MURK

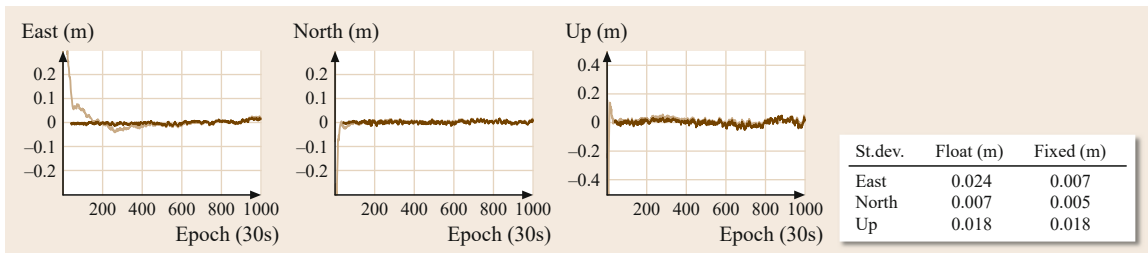


Fig. 26.17 GPS+BDS L1+L2&B1+B2 RTK float (*light*) vs. fixed (*dark*) position errors, as well as standard deviations for 80 km baseline CUT0-MURK

in combination with systematic multipath biases which have a larger impact on the convergence time than the presence of the additional atmospheric parameters in the model, as the convergence time of the short-baseline model is already 4 h. These types of long convergence times for BDS were also observed by [26.70, 71]. GPS combined with BDS has the shortest convergence time, but still almost 3 h. From Fig. 26.17, it follows that mainly for the east component it takes this long time before it has converged to a level below 2 cm.

Integer ambiguity resolution fortunately has a favorable effect. In the GPS-only case, the integers can be fixed after 35 min, while this is 80 min for BDS-only.

Combining GPS and BDS has the shortest time-to-fix, which is just 20 min (Table 26.12). After these times, for all three scenarios the fixed position precision is at the subcentimeter level horizontally and below 3 cm vertically, see Figs. 26.15–26.17. One can see from the tables at the right side of the graphs that the gain due to ambiguity fixing is for the long-baseline cases less than for the short-baseline cases, see the tables corresponding to the short-baseline graphs in Figs. 26.10–26.14. This is because ambiguity resolution needs more time in the long-baseline cases and at the times the ambiguities are finally fixed, the float position precision has already been improved considerably due to the change in receiver-satellite geometry.

26.4 Network RTK

This section deals with the extension of the RTK technique to network RTK. It describes the required processing steps to be performed on the reference station observations, different forms of network RTK realizations, typical correction models, and it discusses PPP-RTK.

26.4.1 From RTK to Network RTK

One significant drawback of single base RTK, as described above, is that the maximum distance between reference station and rover receiver must not exceed 10–20 km in order to be able to rapidly and reliably resolve the carrier-phase ambiguities. This limitation is caused by distance-dependent biases, mainly ionospheric signal refraction but also orbit errors and tropospheric refraction (Sect. 26.1.2). These errors, however, can be accurately modeled using the measurements of an array of GNSS reference stations surrounding the rover site. Thus, the solution to the distance limitation of RTK lies in multibase techniques which became popular under the name network RTK, sometime abbreviated to NRTK. In fact, also network RTK has a distance limitation. This limitation refers to the distances between the reference stations. They should not exceed 100–200 km in order to be able to produce highly accurate real-time correction models of the distance-dependent errors.

Similar to the development of wide-area DGNSS as an extension of local single base DGNSS (Sect. 26.2.1), the network RTK technique enabled the establishment of positioning services which serve larger regions or whole countries by setting up and maintaining networks of reference stations, collecting and preprocessing their observations and distributing observation corrections in real time to RTK users. Only because of the development of network RTK such services became feasible. An area of 100 000 km² requires a network of about 20 reference stations spaced by 75 km if the network RTK technique is used. In comparison, a service using single base RTK with maximum distances to the closest reference station of 7.5 km would require about 900 equally spaced reference receivers for the same area.

26.4.2 Data Processing Methods for Network RTK

Network RTK requires the processing of the reference station observations in order to produce real-time correction models of the distance-dependent errors. Therefore, the reference station observations of a whole

network (or a subnetwork) must be gathered at one data processing site, usually a central processing center. From there, the rover receivers are fed with those reference corrections and model coefficients that they are enabled to perform RTK positioning.

The main data processing steps between gathering the reference station observations and the RTK positioning result are the following (Fig. 26.18).

In the *first processing step* ([1] in Fig. 26.18) ambiguity fixing is performed in the reference station network. This also includes that the undifferenced carrier-phase observations are altered according to the determined double-difference ambiguity values so that afterward all observations are on the same ambiguity level. Only observations with fixed ambiguities can be used for the precise modeling of the distance-dependent errors. This processing step is the crucial part of network RTK because ambiguities of rather long baselines (50–100 km) have to be fixed and they must be fixed in real-time. This network ambiguity resolution differs considerably from common RTK ambiguity resolution since the station coordinates are precisely known. All existing difficulties are caused by observation errors and all a priori information which is able to reduce these errors should be used: IGS predicted satellite ephemerides, ionospheric, and tropospheric corrections based on recent results of the network processing, carrier-phase multipath corrections from the evaluation of the past network data, and antenna phase centre corrections from antenna calibrations.

In the *second processing step* correction model coefficients are estimated ([2] in Fig. 26.18). Several techniques have been developed to model (or interpolate) the distance-dependent biases between reference stations and user receivers. For an overview of potential algorithms, see [26.72] and [26.73]. Ionospheric and orbit biases must be modeled individually for each satellite. Tropospheric corrections, however, may be estimated station by station. It is advantageous to separate the dispersive (ionospheric) biases from the nondispersive biases (orbit and troposphere, sometimes referred to as *geometric*), since ionospheric errors show larger short-term variations as compared to the other distance-dependent biases. Thus, ionospheric corrections must be updated (transmitted to the user) more often, in practice e.g., every 10 s as compared to every 60 s for orbit and tropospheric corrections. Furthermore, in order to be able to correct for small-scale features of ionospheric refraction the modeling area should be kept small, that is, ionospheric correction models may be based on the observations of just three surrounding reference stations. On the other hand, the quality of the geometric

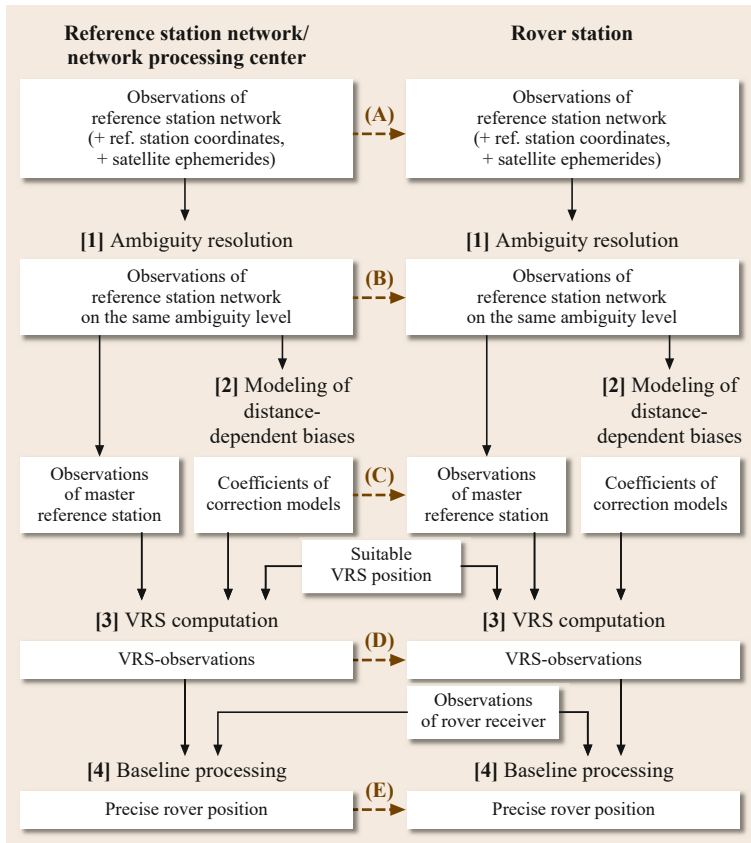


Fig. 26.18 Network RTK processing steps and data transmission options

correction models may improve when a larger number of reference stations is used.

In the *third processing step* of network RTK ([3] in Fig. 26.18), a set of computed reference observations is produced from the real observations of a selected master reference station, often the one closest to the rover receiver, and the precise correction models for distance-dependent biases. Based on the correction models and horizontal coordinate differences between master reference position and approximate rover position, the reference observations are virtually shifted to the rover site. This results in virtual reference station (VRS) observations.

The *fourth and last step* ([4] in Fig. 26.18) consists of the RTK style baseline processing between VRS and rover.

Existing network RTK positioning services run one or more network processing centers where the reference station observations are gathered and where they are preprocessed. The data processing consists at least of real-time ambiguity resolution of the network carrier-phase observations. This goes along with a rigorous quality control of the data. Furthermore, the spatial stability of the reference station antennas is controlled

based on the gathered observations. More processing steps may be performed in the processing center, depending on selected communication links and data formats for transmitting the network data to the user.

Figure 26.18 depicts five ways of how to transfer network information to the user. These approaches differ with respect to allocation of the processing steps either to the network processing center or to the rover. This affects the information content and the data formats of the messages to be transmitted to the rover, and also the selection of appropriate communication channels. Nowadays the most widely used methods are (B) and (D).

Observations of Several Reference Stations

((A) in Fig. 26.18): The rover station receives the observation data streams of several reference stations surrounding his position. After ambiguity resolution of the network data, the rover is able to compute network corrections and even VRS observations to be used for his positioning. The major drawback of this method is related to the ambiguity resolution of the network observations which usually requires an initialization time of several minutes.

Network Observations on Common Ambiguity Level

((B) in Fig. 26.18): Broadcast of observations of a master reference station and observation differences between auxiliary reference stations and master reference station, all being on the same ambiguity level. The user performs the interpolation step on its own and thus obtains network corrections and valuable information on their quality. Then, he is able to compute VRS observations and his position in the baseline mode. This technique is known under the name of master-auxiliary concept (MAC) [26.74, 75].

Coefficients of Correction Model

((C) in Fig. 26.18): Broadcast of observations of a master reference station and coefficients of the correction models of distance-dependent biases. The user applies the correction models to the reference station observation data set according to the coordinate differences between his position and the position of the master reference station, and thus he obtains VRS observations for his site. This technique is often referred to as FKP, an abbreviation of the German term *Flächen-Korrekturparameter* which stands for area correction parameters (Sect. 26.4.3, [26.24, 76]).

Virtual Reference Station (VRS) Observations

((D) in Fig. 26.18): The user sends his approximate position to the central computing facility and by return receives VRS observations to be used for baseline positioning [26.77, 78]. In the early days of network RTK, a major advantage of this technique was that no upgrade to the user equipment software was necessary. But, this method has the drawback that in general no information can be provided on the quality of the interpolation process and thus on the quality of the VRS reference observations.

Precise Rover Position

((E) in Fig. 26.18): The observations of the rover receiver are transmitted to a central processing facility where the baseline processing from VRS to rover is performed. The rover's precise position and quality information may be transmitted back to the user.

Data transfer of reference station observations and broadcast ephemerides to the central processing facility and also of data products from the central processing facility to the user requires adequate communication channels and data formats. The most widely used format for transmission of real-time GNSS observation data and intermediate products is RTCM Standard 10403.2. Positioning results are transmitted using National Marine Electronics Association (NMEA) 1083. In the case of post-processing and transmission of data

files, receiver independent exchange format (RINEX) is the most common open format. It can be used for GNSS observations ((A) and (D) in Fig. 26.18) and broadcast ephemerides, but not for observation differences between stations ((B) in Fig. 26.18) or coefficients of correction models ((C) in Fig. 26.18). More details on data formats are found in Annex .

26.4.3 Network RTK Correction Models

An often used, simple and robust model for network RTK corrections makes use of horizontal gradients of the distance-dependent errors. It is convenient to separate dispersive (ionosphere) from nondispersive (troposphere and orbit) errors. This separation requires dual-frequency carrier-phase observations at the reference stations. The minimum number of reference stations is usually three which surround the service area. The horizontal gradients define a plane correction surface (Fig. 26.19), thus a two-dimensional (2-D) linear interpolation is performed.

These horizontal gradients are usually referred to as FKP. Sets of FKP values are determined for each individual satellite. Such a set consists of four values: namely FKP_{N0} , FKP_{E0} which are the distance-dependent gradients of the nondispersive errors in ppm for north-south and east-west direction, respectively, and FKP_{N1} , FKP_{E1} which are the distance dependent gradients of the dispersive errors scaled to their effects on GPS L1 frequency f_1 in ppm for north-south and east-west direction, respectively. The distance-dependent errors in m for the nondispersive and dispersive components are calculated as

$$\begin{aligned}\delta e_0 &= FKP_{N0} \Delta r_N + FKP_{E0} \Delta r_E, \\ \delta e_1 &= FKP_{N1} \Delta r_N + FKP_{E1} \Delta r_E,\end{aligned}\quad (26.24)$$

where Δr_N and Δr_E are the horizontal coordinate differences in kilometer for the north-south and east-west directions, respectively, of a baseline, for example, between master reference station and rover station.

The distance-dependent errors δe_φ , δe_p for a carrier-phase measurement and a code measurement on frequency f are computed as

$$\begin{aligned}\delta e_\varphi(f) &= \delta e_0 + \frac{f_1^2}{f^2} \delta e_1, \\ \delta e_p(f) &= \delta e_0 - \frac{f_1^2}{f^2} \delta e_1,\end{aligned}\quad (26.25)$$

where f_1 is the GPS L1 frequency.

There are variations of (26.24) and (26.25) which are also in practical use. Differences exist with regard

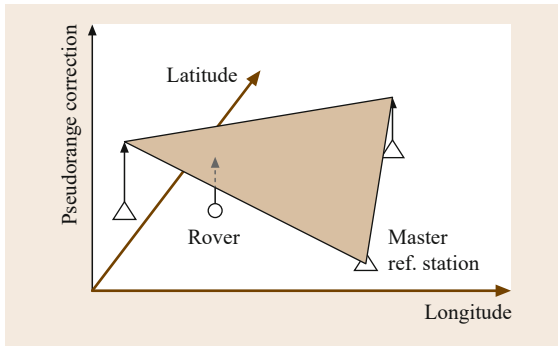


Fig. 26.19 Correction by linear interpolation

to the selected reference frequency to which the dispersive effects are scaled and sometimes the dispersive effects are mapped to the zenith direction [26.36]. This, however, causes no or only negligible changes of the correction effects.

A typical time series of FKP values are shown in Fig. 26.20. Each line connects the model coefficients of a specific satellite. The FKP for the dispersive effects (lower panels) are often larger in size and show more short-term variations. This emphasizes that one should consider updating them more frequently as those values which correct nondispersive effects, for example, every 10 s as compared to every 60 s (*upper panels of Fig. 26.20*).

26.4.4 Refined Virtual Reference Stations

Computing VRS observations involves the shift of the master reference observations to the selected position of the VRS. This shift actually consist of two different processes: one performs the geometrical shift of the observations mainly considering the changing distances to the satellites, the other applies corrections for the distance-dependent biases according to the latitudinal and longitudinal differences between master reference station position and VRS position. In general, the same VRS position is used for both processes. Only then, the VRS is able to resemble real observation data collected at this site.

There are reasons to refine this concept and to distinguish between two different positions of a VRS: the geometric position and the position used to apply the correction models (Fig. 26.21). The latter position should be as close as possible to the actual rover position. The geometric position of the VRS may be a different one.

Experience shows that parts of the ionospheric errors can be of very small-scale size, so that they are not completely captured in a regional network of reference stations. Thus, the baseline from the VRS to the

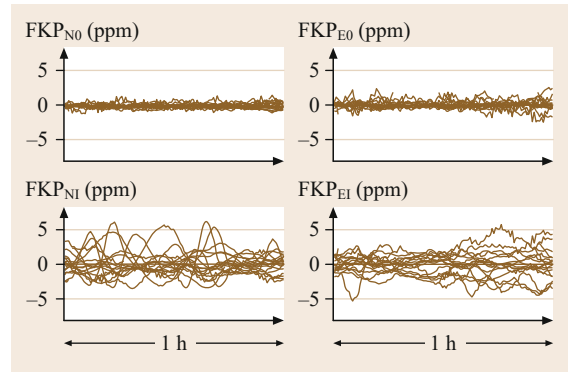


Fig. 26.20 Examples of FKP values

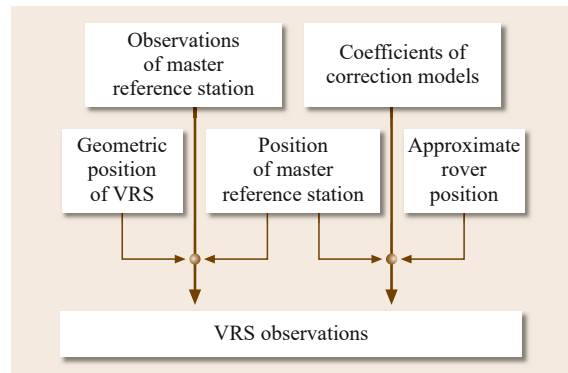


Fig. 26.21 Modification to the generation of VRS observations

rover often contains more remaining ionospheric errors than a short baseline between a real reference station and the rover would do. As a consequence, the baseline processing between VRS and rover must not assume a complete elimination of ionospheric effects which would be correct for a very short baseline between two real receivers. It should assume that the residual ionospheric effects are as large as those of a baseline of a few or even more kilometers depending on the actual ionospheric conditions, on the distances between the real reference stations, and on the modeling approach.

In order to communicate this information to the user, a different form of virtual reference station is created, a so-called PRS or i-MAX [26.79]. The geometric position of the VRS is selected in such a way that the baseline length to the rover receiver is able to reflect the size of the remaining ionospheric errors which can then be handled by an appropriate stochastic modeling approach.

Another kind of refined VRS is suggested for larger scale kinematic applications. A moving rover requires changing positions of the virtual reference sta-

tion. Since baseline processing software do not accept a moving reference station, a so-called semikinematic VRS [26.80] is produced. The corrections of the distance-dependent errors are applied according to the changing positions of the moving rover. At the same time, a fixed position of the VRS is used to imitate best a reference station.

Such refined VRS are produced for individual rovers and they must only be used for baseline processing with those rover receiver.

26.4.5 From Network RTK to PPP-RTK

In recent years, precise point positioning (PPP, Chap. 25) without or with ambiguity fixing started to compete with the differential positioning technique RTK. One of the main drawbacks of PPP is the fairly long convergence time of several to many minutes which is needed to obtain accurate ambiguity estimates or a reliable ambiguity fixing. With RTK or network

RTK the convergence times usually do not exceed 1 or 2 min and the obtainable positioning accuracies exceed those of PPP.

The main reason for these differences lies in the correction or elimination by differencing of ionospheric and tropospheric errors. Local (RTK) or regional (network RTK) reference stations can be considered as atmospheric sensors whose information is able to speed up ambiguity fixing. This leads to PPP-RTK where faster PPP ambiguity fixing and also higher accurate positioning results are obtained by including atmospheric corrections based on observation data of local or regional GNSS reference stations. The overall performance of PPP-RTK is similar to the one of network RTK [26.81–83].

Acknowledgments. We would like to thank Dr Amir Khodabandeh of the GNSS Research Centre at Curtin University for his help with the writing of the GLONASS RTK section.

References

- 26.1 Y. Georgiadou, K.D. Doucet: The issue of selective availability, *GPS World* **1**(5), 53–56 (1990)
- 26.2 F. van Graas, M.S. Braasch: Selective availability. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B. Parkinson, J.J. Spilker Jr. (AIAA, Washington 1995) pp. 601–621
- 26.3 J.K. Gupta, L. Singh: Long term ionospheric electron content variations over Delhi, *Ann. Geophysicae* **18**, 1635–1644 (2001)
- 26.4 S. Skone, S.M. Shrestha: Limitations in DGPS positioning accuracies at low latitudes during solar maximum, *Geophys. Res. Lett.* **29**(10), 81/1–81/4 (2002)
- 26.5 R. Warnant: Influence of the ionospheric refraction on the repeatability of distances computed by GPS, *Proc. ION GPS-97*, Kansas City (ION, Virginia 1997) pp. 217–224
- 26.6 A.J. Coster, M.M. Pratt, B.P. Burke, P.N. Misra: Characterization of atmospheric propagation errors for DGPS, *Proc. ION AM-98*, Denver (ION, Virginia 1998) pp. 327–336
- 26.7 H.B. Vo, J.C. Foster: Quantitative investigation of ionospheric density gradients at mid latitudes, *J. Geophys. Res.* **106**, 21555–21563 (2001)
- 26.8 L. Wanninger: Effects of the equatorial ionosphere on GPS, *GPS World* **7**(4), 48–54 (1993)
- 26.9 S. Skone, M. El-Gizawy, S.M. Shrestha: Limitations in GPS positioning accuracies and receiver tracking performance during solar maximum, *Proc. Kinem. Syst. Geod., Geom. Navig. (KIS2001)*, Banff (University of Calgary, Calgary 2001) pp. 129–143
- 26.10 J. Saastamoinen: Atmospheric correction for the troposphere and stratosphere in radio ranging of satellites, *Geophys. Monogr. Ser.* **15**, 247–251 (1972)
- 26.11 F. Kleijer: Troposphere Modeling and Filtering for Precise GPS Leveling, Ph.D. Thesis (Netherlands Geodetic Commission, Delft 2004)
- 26.12 H.B. Baby, P. Gole, J. Lavergnat: A model for the tropospheric excess path length of radio waves from surface meteorological measurements, *Radio Sci.* **23**(6), 1023–1038 (1988)
- 26.13 J.J. Spilker Jr.: Tropospheric effects on GPS. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B. Parkinson, J.J. Spilker Jr. (AIAA, Washington 1995) pp. 517–546
- 26.14 P. Bona, C. Tiberius: An experimental comparison of noise characteristics of seven high-end dual frequency GPS receiver sets, *Proc. IEEE Position Location Navig. Symp.*, San Diego (2000) pp. 237–244
- 26.15 S.H. Byun, G.A. Hajj, L.E. Young: GPS signal multipath: A software simulator, *GPS World* **13**(7), 40–49 (2002)
- 26.16 J. Raquet, G. Lachapelle: Determination and reduction of GPS reference station multipath using multiple receivers, *Proc. ION GPS-96*, Kansas City (ION, Virginia 1996) pp. 673–681
- 26.17 M.S. Braasch, A.J. van Dierendonck: GPS receiver architectures and measurements, *Proc. IEEE* **87**(1), 48–64 (1999)
- 26.18 P. Bona: Accuracy of GPS phase and code observations in practice, *Acta Geod. Geophys. Hung.* **35**(4), 433–451 (2000)
- 26.19 P.F. de Bakker, C.C.J.M. Tiberius, H. van der Marel, R.J.P. van Bree: Short and zero baseline analysis of GPS L1 C/A, L5Q, GIOVE E1B and E5aQ signals, *GPS Solutions* **16**(1), 53–64 (2012)
- 26.20 T.H. Diessongo, T. Schüller, S. Junker: Precise position determination using a Galileo E5 single-fre-

- quency receiver, *GPS Solutions* **7**(4), 230–240 (2013)
- 26.21 D. Odijk, P.J.G. Teunissen, A. Khodabandeh: Galileo IOV RTK positioning: Standalone and combined with GPS, *Surv. Rev.* **46**(337), 267–277 (2014)
- 26.22 C. Cai, C. He, R. Santerre, L. Pan, X. Cui, J. Zhu: A comparative analysis of measurement noise and multipath for four constellations: GPS, BeiDou, GLONASS and Galileo, *Surv. Rev.* **48**(349), 287–295 (2016)
- 26.23 B. Parkinson, P. Enge: Differential GPS. In: *Global Positioning System: Theory and Applications*, Vol. 2, ed. by B. Parkinson, J.J. Spilker Jr. (AIAA, Washington 1995) pp. 3–50
- 26.24 G. Wübbena, A. Bagge, G. Seeber, V. Böder, P. Hankemeier: Reducing distance dependent errors for real-time precise DGPS applications by establishing stations networks, *Proc. ION GPS-96*, Kansas City (ION, Virginia 1996) pp. 1845–1852
- 26.25 P.J.G. Teunissen: Differential GPS: Concepts and quality control, *Proc. NIN Workshop Glob. Position. Syst.*, Amsterdam (Netherlands Institute of Navigation, Delft 1991) pp. 1–46
- 26.26 C. Kee, B.W. Parkinson: Wide area differential GPS (WADGPS): Future navigation system, *IEEE Trans. Aerosp. Electron. Syst.* **32**(2), 795–808 (1996)
- 26.27 H. Rho, R.B. Langley: Dual-frequency GPS precise point positioning with WADGPS corrections, *Proc. ION GNSS 2005*, Long Beach (ION, Virginia 2005) pp. 1470–1482
- 26.28 J.A. Klobuchar: Ionospheric time-delay algorithm for single-frequency GPS users, *IEEE Trans. Aerosp. Electron. Syst.* **23**(3), 325–331 (1986)
- 26.29 X. Wu, X. Hu, G. Wang, H. Zhong, C. Tang: Evaluation of COMPASS ionospheric model in GNSS positioning, *Adv. Space Res.* **51**, 959–968 (2013)
- 26.30 G. di Giovanni, S.M. Radicella: An analytical model of the electron density profile in the ionosphere, *Adv. Space Res.* **10**(11), 27–30 (1990)
- 26.31 Y. Yuan, X. Huo, J. Ou, K. Zhang, Y. Chai, D. Wen, R. Grenfell: Refining the Klobuchar ionospheric coefficients based on GPS observations, *IEEE Trans. Aerosp. Electron. Syst.* **44**(4), 1498–1510 (2008)
- 26.32 A. Angrisano, S. Gaglione, C. Gioia, M. Massaro, U. Robustelli: Assessment of NeQuick ionospheric model for Galileo single-frequency users, *Acta Geophysica* **61**(6), 1457–1476 (2013)
- 26.33 J. Boehm, R. Heinkelmann, H. Schuh: Short note: A global model of pressure and temperature for geodetic applications, *J. Geod.* **81**(10), 679–683 (2007)
- 26.34 D.B. Wolfe, C.L. Judy, E.J. Haukka, D.J. Godfrey: Implementing and engineering an NDGPS network in the United States, *Proc. ION GPS 2000*, Salt Lake City (ION, Virginia 2000) pp. 1254–1263
- 26.35 A. Cameron, T. Reynolds: NDGPS loses interior, keeps coast, *GPS World* **27**(8), 9 (2016)
- 26.36 Radio Technical Commission for Maritime Services: RTCM Standard 10403.2 Differential GNSS Services, Version 3 with Amendment 5 (RTCM, Arlington 2013)
- 26.37 G. Weber, D. Dettmering, H. Gebhard, R. Kalafus: Networked transport of RTCM via internet protocol (Ntrip) – IP-streaming for real-time GNSS applications, *Proc. ION GPS 2005*, Long Beach (ION, Virginia 2005) pp. 2243–2247
- 26.38 B. Park, J. Kim, C. Kee: RRC unnecessary for DGPS messages, *IEEE Trans. Aerosp. Electron. Syst.* **42**(3), 1149–1160 (2006)
- 26.39 D. Dettmering, G. Weber: The EUREF-IP Ntrip broadcaster: Real-time GNSS data for Europe, *Proc. IGS Workshop*, Bern (2004)
- 26.40 P.J.G. Teunissen: An integrity and quality control procedure for use in multi sensor integration, *Proc. ION GPS 1990*, Colorado Springs (ION, Virginia 1990) pp. 513–522
- 26.41 P.J.G. Teunissen: GPS double difference statistics: With and without using satellite geometry, *J. Geod.* **71**(3), 137–148 (1997)
- 26.42 B.W. Remondi: Using the Global Positioning System (GPS) Phase Observable for Relative Geodesy: Modeling, Processing and Results, Ph.D. Thesis (University of Texas, Austin 1984)
- 26.43 Y. Bock, R.I. Abbot, C.C. Counselman, S.A. Gourevitch, R.W. King: Establishment of three-dimensional geodetic control by interferometry with the global positioning system, *J. Geophys. Res.* **90**(B9), 7689–7703 (1985)
- 26.44 B.W. Remondi: Performing centimeter accuracy relative surveys in seconds using carrier phase, *Proc. 1st Int. Symp. Precise Position. Glob. Position. Syst. (NOAA)*, Rockville (National Geodetic Information Center, NOAA, Rockville 1985) pp. 789–797
- 26.45 C.C. Goad: Precise positioning with the global positioning system, *Proc. 3rd Int. Symp. Inertial Technol. Surv. Geod.* (1986) pp. 745–756
- 26.46 M.E. Cannon: High accuracy GPS semikinematic positioning: Modeling and results, *Navigation* **37**(1), 53–64 (1990)
- 26.47 Y. Kubo, Y. Muto, S. Kitao, C. Uratan, S. Sugimoto: Ambiguity resolution for dual frequency carrier phase kinematic GPS, *Proc. IEEE TENCON* (2004) pp. 661–664
- 26.48 T. Takasu, A. Yasuda: Kalman-filter-based integer ambiguity resolution strategy for long-baseline RTK with ionosphere and troposphere estimation, *Proc. ION GNSS 2010*, Portland (ION, Virginia 2010) pp. 161–171
- 26.49 P.J. de Jonge: A Processing Strategy for the Application of the GPS in Networks, Ph.D. Thesis (Netherlands Geodetic Commission, Delft 1998)
- 26.50 D. Odijk: Fast Precise GPS Positioning in the Presence of Ionospheric Delays, Ph.D. Thesis (Netherlands Geodetic Commission, Delft 2002)
- 26.51 M. Pratt, B. Burke, P. Misra: Single-epoch integer ambiguity resolution with GPS-GLONASS L1-L2 data, *Proc. ION GPS 1998*, Nashville (ION, Virginia 1998) pp. 389–398
- 26.52 N. Reussner, L. Wanninger: GLONASS inter-frequency biases and their effects on RTK and PPP carrier-phase ambiguity resolution, *Proc. ION GNSS 2011*, Portland (ION, Virginia 2011) pp. 712–716
- 26.53 F. Takac: GLONASS inter-frequency biases and ambiguity resolution, *Inside GNSS* **4**(2), 24–28 (2009)
- 26.54 L. Wanninger, S. Wallstab-Freitag: Combined processing of GPS, GLONASS and SBAS code phase and

- carrier phase measurements, Proc. ION GNSS 2007, Fort Worth (ION, Virginia 2007) pp. 866–875
- 26.55 L. Wanninger: Carrier-phase inter-frequency biases of GLONASS receivers, *J. Geod.* **86**(2), 139–148 (2012)
- 26.56 H. Yamada, T. Takasu, N. Kubo, A. Yasuda: Evaluation and calibration of receiver inter-channel biases for RTK-GPS/GLONASS, Proc. ION GNSS 2010, Portland (ION, Virginia 2010) pp. 1580–1587
- 26.57 J. Wang, C. Rizos, M.P. Stewart, A. Leick: GPS and GLONASS integration: Modeling and ambiguity resolution issues, *GPS Solutions* **5**(1), 55–64 (2001)
- 26.58 A. Leick, J. Li, Q. Beser, G. Mader: Processing GLONASS carrier phase observations – Theory and first experience, Proc. ION GPS 1995, Palm Springs (ION, Virginia 1995) pp. 1041–1047
- 26.59 S. Banville, P. Collins, F. Lahaye: GLONASS ambiguity resolution of mixed receiver types without external calibration, *GPS Solutions* **17**(3), 275–282 (2013)
- 26.60 D. Odijk, P.J.G. Teunissen, L. Huisman: First results of mixed GPS+GLOVE single-frequency RTK in Australia, *J. Spatial Sci.* **57**(1), 3–18 (2012)
- 26.61 R. Odolinski, P.J.G. Teunissen, D. Odijk: Combined GPS+BDS+Galileo+QZSS for long baseline RTK positioning, Proc. ION GNSS+ 2014, Tampa (ION, Virginia 2014) pp. 2326–2340
- 26.62 O. Julien, P. Alves, M.E. Cannon, W. Zhang: A tightly coupled GPS/GALILEO combination for improved ambiguity resolution, Proc. ENC/GNSS, Graz (Austrian Institute of Navigation (OVN), Graz 2003)
- 26.63 D. Odijk, P.J.G. Teunissen: Characterization of between-receiver GPS-Galileo inter-system biases and their effect on mixed ambiguity resolution, *GPS Solutions* **17**(4), 521–533 (2013)
- 26.64 J. Paziewski, P. Wielgosz: Accounting for Galileo-GPS inter-system biases in precise satellite positioning, *J. Geod.* **89**(1), 81–93 (2015)
- 26.65 N. Nadarajah, P.J.G. Teunissen, N. Raziq: BeiDou inter-satellite-type bias evaluation and calibration for mixed receiver attitude determination, *Sensors* **13**, 9435–9463 (2013)
- 26.66 N. Nadarajah, P.J.G. Teunissen, J.-M. Sleewaegen, O. Montenbruck: The mixed-receiver BeiDou inter-satellite-type bias and its impact on RTK positioning, *GPS Solutions* **19**(3), 357–368 (2015)
- 26.67 G. Wang, K. de Jong, Q. Zhao, Z. Hu, J. Guo: Multipath analysis of code measurements for BeiDou geostationary satellites, *GPS Solutions* **19**(1), 129–139 (2015)
- 26.68 P.J.G. Teunissen, R. Odolinski, D. Odijk: Instantaneous BeiDou+GPS RTK positioning with high cut-off elevation angles, *J. Geod.* **88**(4), 335–350 (2014)
- 26.69 N. Nadarajah, P.J.G. Teunissen: Instantaneous GPS/Galileo/QZSS/SBAS attitude determination: A single-frequency (L1/E1) robustness analysis under constrained environments, *Navigation* **61**(1), 65–75 (2014)
- 26.70 M. Wang, H. Cai, Z. Pan: BDS/GPS relative positioning for long baseline with undifferenced observations, *Adv. Space Res.* **55**, 113–124 (2014)
- 26.71 X. Zhang, X. He: Performance analysis of triple-frequency ambiguity resolution with BeiDou observations, *GPS Solutions* **20**(2), 269–281 (2016)
- 26.72 G. Fotopoulos, M.E. Cannon: An overview of multiple-reference station methods for cm-level positioning, *GPS Solutions* **4**, 1–10 (2001)
- 26.73 L. Dai, S. Han, J. Wang, C. Rizos: A study on GPS/GLONASS multiple reference station techniques for precise real-time carrier phase based positioning, Proc. ION GPS 2001, Salt Lake City (ION, Virginia 2001) pp. 392–403
- 26.74 H.-J. Euler, C.R. Keenan, B.E. Zebhauser, G. Wübbena: Study of a simplified approach in utilizing information from permanent reference station arrays, Proc. ION GPS 2001, Salt Lake City (ION, Virginia 2001) pp. 379–391
- 26.75 H.-J. Euler, S. Seeger, O. Zelzer, F. Takac, B.E. Zebhauser: Improvement of positioning performance using standardized network RTK messages, Proc. ION NTM 2004, San Diego (ION, Virginia 2004) pp. 453–461
- 26.76 G. Wübbena, A. Bagge: Neuere Entwicklungen zu GNSS-RTK für optimierte Genauigkeit, Zuverlässigkeit und Verfügbarkeit: Referenzstationen und Multistations-RTK-Lösungen, proc. 46th DVW-Fortbildungsseminar: GPS-Praxis und Trends '97, DVW-Schriftenreihe 35/1999, Frankfurt (1999) pp. 73–92
- 26.77 L. Wanninger: Real-time differential GPS error modelling in regional reference station networks. In: *International Association of Geodesy Symposia*, Vol. 118, (Springer, Berlin 1997) pp. 86–92
- 26.78 U. Vollath, A. Buecherl, H. Landau, C. Pagels, B. Wagner: Multi-base RTK positioning using virtual reference stations, Proc. ION GPS 2000, Salt Lake City (ION, Virginia 2000) pp. 123–131
- 26.79 F. Takac, O. Zelzer: The relationship between network RTK solutions MAC, VRS, PRS, FKP and i-MAX, Proc. ION GPS 2008, Savannah (ION, Virginia 2008) pp. 348–355
- 26.80 L. Wanninger: Real-time differential GPS error modelling in regional reference station networks, Proc. ION GPS 2002, Portland (ION, Virginia 2002) pp. 1400–1407
- 26.81 G. Wübbena, M. Schmitz, A. Bagge: PPP-RTK: Precise point positioning using state-space representation in RTK networks, Proc. ION GNSS 2005, Long Beach (ION, Virginia 2005) pp. 2584–2594
- 26.82 P.J.G. Teunissen, D. Odijk, B. Zhang: PPP-RTK: Results of CORS network-based PPP with integer ambiguity resolution, *J. Aeronaut. Astronaut. Aviat. Ser. A* **42**(4), 223–230 (2010)
- 26.83 X. Li, M. Ge, J. Douša, J. Wickert: Real-time precise point positioning regional augmentation for large GPS reference networks, *GPS Solutions* **18**(1), 61–71 (2014)

Attitude Dete

27. Attitude Determination

Gabriele Giorgi

Attitude estimation is the process of determining the spatial orientation of an object. A system formed by multiple Global Navigation Satellite System (GNSS) antennas placed at known relative positions acts as an attitude sensor. This chapter provides an overview of practical applications of GNSS-based attitude determination, gives the principles of attitude representation and estimation, and reviews a constrained ambiguity resolution method to reliably fix the carrier-phase integer ambiguities and obtain precise attitude estimations.

27.1	Six Degrees of Freedom	781	27.3.2	Orthogonal Procrustes Problem	788
27.2	Attitude Parameterization	784	27.3.3	Weighted Orthogonal Procrustes Problem	789
27.2.1	The Space of Rotations	784	27.3.4	Attitude Estimation with Fully Populated Weight Matrix	789
27.2.2	Parameterization of the Rotation Matrix	784	27.3.5	On the Precision of Attitude Estimation	790
27.3	Attitude Estimation from Baseline Observations	787	27.4	The GNSS Attitude Model	790
27.3.1	Estimation of the Orthonormal Matrix of Rotations	787	27.4.1	Potential Model Errors and Misspecification	791
			27.4.2	Resolution of the GNSS Attitude Model	792
			27.4.3	The GNSS Ambiguity and Attitude Estimation	793
			27.4.4	The Quality of Ambiguity and Attitude Estimations	795
			27.5	Applications	798
			27.5.1	Space Operations	798
			27.5.2	Aeronautics Applications	800
			27.5.3	Marine Navigation	802
			27.5.4	Land Applications	803
			27.6	An Overview of GNSS/INS Sensor Fusion for Attitude Determination	804
			References		806

27.1 Six Degrees of Freedom

The pose of a rigid body in a three dimensional space can be described by six independent parameters, of which three refer to the absolute position of a body reference point and three describe the body orientation. Whereas positioning deals with the estimation of the absolute location, attitude estimation is the process of determining the orientation of the body with respect to a reference frame. The combination of absolute position and orientation enables a complete static characterization of a rigid body in space, as shown in Fig. 27.1.

Several attitude sensors are available to obtain the spatial orientation of a body. These can be classified in two distinct categories: relative and absolute sensors. Relative attitude sensors detect changes in the body dynamics by exploiting internal devices, and keep track of rotational accelerations induced by the body mo-

tion, generating an output proportional to the magnitude of the rotations. The absolute orientation of the body is then obtained by continuously propagating a known initial state. Due to the accumulation of measurement errors, the integration process causes estimation biases that tend to increase over time, thus requiring periodic recalibrations. An example of a relative attitude sensor is the gyroscope, which reacts to variations of the body orientation by opposing a measurable gyroscopic resistance force.

Absolute attitude sensors detect the body orientation by exploiting an external source, independent of the body whose attitude has to be determined. Examples of absolute sensors are star trackers, which relate the body orientation with respect to the direction pointing to a group of selected stars; magnetometers, which sense the direction of an external magnetic field; and

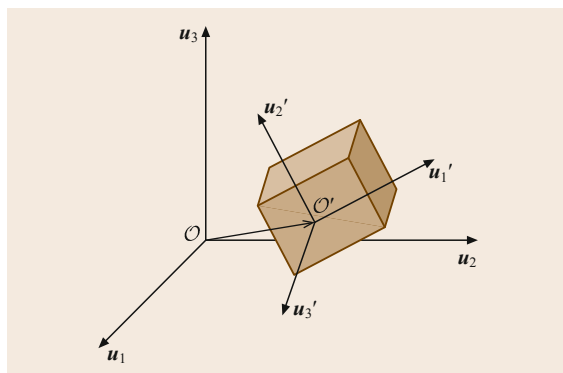


Fig. 27.1 The pose of a rigid body in space is completely characterized by the position of a reference point O' and the rotation of a frame integral with the body with respect to a reference frame

horizon sensors, which track the angular direction(s) of the line(s) of sight to near-horizon regions.

Global Navigation Satellite System (GNSS) antennas can be employed as absolute attitude sensors. A coarse indication of the orientation of an antenna with respect to the line-of-sight vector to the tracked satellite can be obtained by comparing the received signal strength against the antenna radiation pattern. Multiple error sources may alter the received power levels, such as multipath and atmospheric disturbances, and rotations about the antenna boresight axis are not observable with symmetrical antenna gain patterns. Typical accuracies are in the order of ten degrees, with error peaks reaching several tens of degrees [27.1].

A system formed by two or more GNSS antennas is better suited to provide orientation estimates. The idea of using GNSS signals to obtain attitude information dates back to 1976, when the concept of supplying a spacecraft in a low Earth orbit with two Global

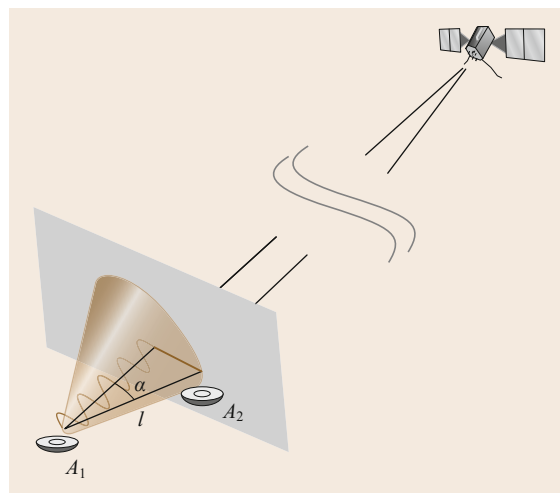


Fig. 27.2 The differential range measurement at two antennas tracking the same satellite provides information about the potential direction of the baseline vector

Positioning System (GPS) antennas placed at known distance was first proposed [27.2]. The initial concept was then refined, principally by including carrier-phase measurements to improve the ranging capabilities of the GPS-based differential positioning [27.3, 4]. After initial successful tests on static platforms [27.5, 6], several tests on vessels [27.7] and aircrafts [27.8] demonstrated the viability of GPS single-baseline heading estimation, whereas the full capabilities of a GPS antenna array for three-axis attitude determination was first proven on a real-time flight test aboard a DC-3 aircraft in 1991 [27.9]. Over the course of the last two decades constant improvements in the areas of fast initialization procedures, reliable on-the-fly carrier-phase ambiguity estimation, and enhanced accuracy in real-time dynamic environments have been made, and GNSS-based

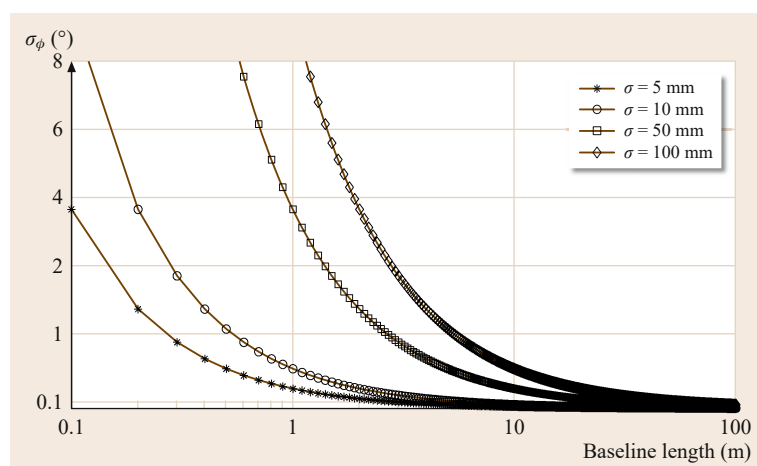


Fig. 27.3 Coarse prediction of attitude estimation precision (σ_ϕ) as function of the baseline length and differential range measurement precision σ

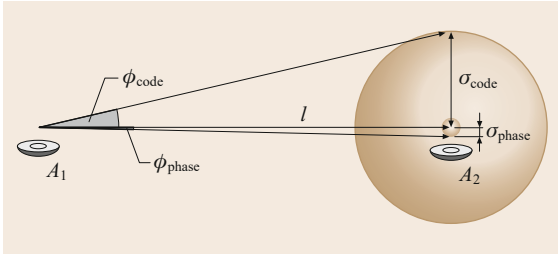


Fig. 27.4 GNSS carrier-phase measurements provide much more precise differential positioning than pseudorange (code) measurements, thus enabling highly precise angular estimates

attitude determination systems have found a firm place in the spectrum of attitude sensors.

GNSS signals are processed to obtain directional information by exploiting the interferometric principle: the difference between ranging measurements at two antennas tracking the same navigation satellite equals to the projection of the baseline vector – that is, the vectorial distance between the two antennas – onto the line-of-sight direction to the common satellite (Fig. 27.2). Each differential range measurement corresponds to a locus of potential baseline vector directions, which span a conical region of aperture $\alpha = \arccos(r/l)$, with r the differential range and l the baseline length. The baseline vector can then be estimated when three or more satellites are being simultaneously tracked. A single baseline enables direction measurements – for example the heading and pitch of a vehicle – whereas two or more noncollinear baselines are needed to perform full three-axis attitude determination.

The quality of angular estimation from differential range measurements depends on two factors: the baseline length and the differential ranging error. A simple rule of thumb to predict the angular estimation error σ_ϕ from a given combination of baseline length l and differential ranging measurement error σ is

$$\sigma_\phi = \frac{\sigma}{l} . \quad (27.1)$$

This relationship is shown in Fig. 27.3: subdegree accuracies are obtained by employing large antenna separations or reducing the differential measurement error.

The latter can be minimized in the context of GNSS attitude sensors by exploiting carrier-phase measurements. Whereas pseudorange differential measurement error is at decimeter level at best, the signal carrier phase can be measured at a fraction of the wavelength, enabling differential positioning with subcentimeter precision (Fig. 27.4). However, each carrier-phase measurement is ambiguous by an unknown integer number of cycles: only the fractional part of the differential signal phase can be observed (Fig. 27.5). Carrier-phase integer ambiguity resolution is the process of resolving the cycle ambiguities to their correct integer values (Chap. 23). After the integer ambiguities are removed, the differential carrier-phase observations act as very precise ranging measurements, yielding precise baseline estimations which are then converted to an estimate of the orientation angles. Typically, a GNSS carrier-phase-based attitude determination sensor enables orientation estimation with subdegree precision for baselines of about 1 m length and with precision higher than 0.1° for baselines spanning few meters.

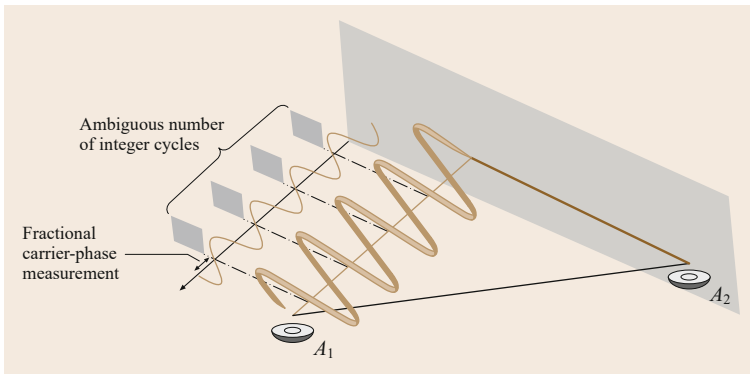


Fig. 27.5 The differential carrier-phase measurement is ambiguous by an integer number of cycles, since only the fractional part of the signal phase can be observed. Integer ambiguity resolution is the key to precise attitude estimates

27.2 Attitude Parameterization

Prior to describing how to perform attitude determination with GNSS signals, a review of attitude representation methods is here given. The orientation of a body is described by defining the transformation matrix that rotates a coordinate system attached to the body onto a reference system of choice. Several alternatives for representing the mutual orientation of two frames are available. Each parameterizes the rotation by employing a small set of *attitude parameters*.

This section gives an introduction to the properties of the transformation matrix under the rigid-body hypothesis, and briefly reviews several alternatives to efficiently represent the rotation in terms of attitude parameters.

27.2.1 The Space of Rotations

An orthogonal frame \mathcal{F} in the vector space \mathbb{R}^3 is defined by a basis, that is, a triplet of vectors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ that satisfy the following two conditions

$$\mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}, \quad (27.2)$$

$$\mathbf{u}_1 \times \mathbf{u}_2 = \mathbf{u}_3, \quad (27.3)$$

with δ_{ij} the Kronecker delta. The elements of the basis are thus orthogonal vectors of norm one, and define a right-handed coordinate system.

Any vector $\mathbf{x} \in \mathbb{R}^3$ can be expressed in terms of its projection onto the basis of \mathcal{F}

$$\mathbf{x} = (\mathbf{x}^\top \mathbf{u}_1) \mathbf{u}_1 + (\mathbf{x}^\top \mathbf{u}_2) \mathbf{u}_2 + (\mathbf{x}^\top \mathbf{u}_3) \mathbf{u}_3. \quad (27.4)$$

The components of vector \mathbf{x} in any two orthogonal frames \mathcal{F} and \mathcal{F}' sharing a common origin are related by the linear transformation

$$\mathbf{x}_{\mathcal{F}'} = \mathbf{R} \mathbf{x}_{\mathcal{F}}. \quad (27.5)$$

Matrix \mathbf{R} defines the transformation that rotates the elements of the basis of \mathcal{F} onto the elements of the basis of \mathcal{F}' . This transformation must preserve both the lengths and the angles between any two vectors \mathbf{x} and \mathbf{y} in \mathbb{R}^3

$$\mathbf{x}_{\mathcal{F}'}^\top \mathbf{y}_{\mathcal{F}'} = \mathbf{x}_{\mathcal{F}}^\top \mathbf{R}^\top \mathbf{R} \mathbf{y}_{\mathcal{F}} \implies \mathbf{R}^\top \mathbf{R} = \mathbf{I}_3. \quad (27.6)$$

Thus, matrix \mathbf{R} is orthonormal, that is, the vector columns \mathbf{r}_i of \mathbf{R} are unit vectors, pair-wise orthogonal

$$\mathbf{r}_i^\top \mathbf{r}_j = \delta_{ij}. \quad (27.7)$$

Condition (27.7) guarantees the invariance of the scalar product with respect to the linear transformation (27.5),

whereas the vector product is invariant under rotations about the axis defined by $\mathbf{x}_{\mathcal{F}'} \times \mathbf{y}_{\mathcal{F}'}$ only for rotation matrices with positive determinant

$$\mathbf{x}_{\mathcal{F}'} \times \mathbf{y}_{\mathcal{F}'} = \det(\mathbf{R}) \mathbf{R} (\mathbf{x}_{\mathcal{F}} \times \mathbf{y}_{\mathcal{F}}). \quad (27.8)$$

An orthonormal matrix with positive determinant is named an attitude matrix, or equivalently a matrix of rotations. An orthonormal matrix with negative determinant describes a combination of a rotation and a reflection about one or more axes, and for this reason it does not describe transformations of real rigid bodies. The group of orthonormal matrices with positive determinant is named a special group $SO(3)$ [27.10].

Due to the orthonormality constraint, in the three-dimensional space the matrix of rotations \mathbf{R} can be parameterized with a minimum of three independent variables.

27.2.2 Parameterization of the Rotation Matrix

The transformation between the basis of two orthonormal frames \mathcal{F} and \mathcal{F}' reads

$$\mathbf{u}_{i,\mathcal{F}'} = \sum_{j=1}^3 r_{ij} \mathbf{u}_{j,\mathcal{F}}, \quad i = 1, 2, 3, \quad (27.9)$$

where r_{ij} denotes the element of matrix \mathbf{R} . Applying relation (27.9) to the scalar product between vectors of the two bases yields

$$\mathbf{u}_{i,\mathcal{F}'}^\top \mathbf{u}_{j,\mathcal{F}} = r_{ij}. \quad (27.10)$$

The scalar r_{ij} is the cosine of the angle between vectors $\mathbf{u}_{i,\mathcal{F}'}$ and $\mathbf{u}_{j,\mathcal{F}}$. The elements of the attitude matrix then corresponds to the nine direction cosines of the angles formed by the three vectors of basis \mathcal{F} and the three vectors of basis \mathcal{F}' (Fig. 27.6)

$$\mathbf{R} = \begin{bmatrix} \mathbf{u}_{1,\mathcal{F}'}^\top \mathbf{u}_{1,\mathcal{F}} & \mathbf{u}_{1,\mathcal{F}'}^\top \mathbf{u}_{2,\mathcal{F}} & \mathbf{u}_{1,\mathcal{F}'}^\top \mathbf{u}_{3,\mathcal{F}} \\ \mathbf{u}_{2,\mathcal{F}'}^\top \mathbf{u}_{1,\mathcal{F}} & \mathbf{u}_{2,\mathcal{F}'}^\top \mathbf{u}_{2,\mathcal{F}} & \mathbf{u}_{2,\mathcal{F}'}^\top \mathbf{u}_{3,\mathcal{F}} \\ \mathbf{u}_{3,\mathcal{F}'}^\top \mathbf{u}_{1,\mathcal{F}} & \mathbf{u}_{3,\mathcal{F}'}^\top \mathbf{u}_{2,\mathcal{F}} & \mathbf{u}_{3,\mathcal{F}'}^\top \mathbf{u}_{3,\mathcal{F}} \end{bmatrix}. \quad (27.11)$$

By employing nine direction cosines, the mutual orientation of two frames is described with six redundant parameters. A more efficient approach is obtained by exploiting Euler's rotation theorem [27.11]: any rotation of a rigid body in a three-dimensional space can be described with a single rotation with angle ϕ about an axis \mathbf{n} . The latter is identified by the three direction cosines that it forms with the basis of frame \mathcal{F}

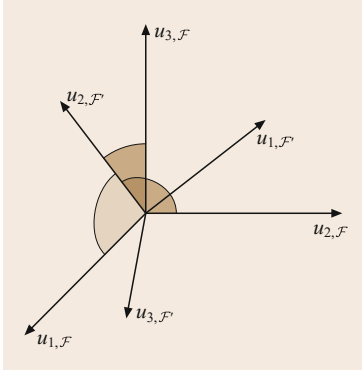


Fig. 27.6 The nine angles formed by the axes of frame \mathcal{F} and the axes of frame \mathcal{F}' unambiguously define the mutual orientation of the two frames. Here, the three angles formed by axis $u_{2,\mathcal{F}'}$ with respect to the three axes of frame \mathcal{F} are shown

(Fig. 27.7). The Euler rotation formula gives the parameterization of the attitude matrix in terms of angle ϕ and the elements of vector \mathbf{n}

$$\mathbf{R}(\mathbf{n}, \phi) = C_\phi \mathbf{I}_3 + (1 - C_\phi) \mathbf{n} \mathbf{n}^\top - S_\phi [\mathbf{n}^+] , \quad (27.12)$$

where $C_\phi = \cos(\phi)$ and $S_\phi = \sin(\phi)$. The skew-symmetric matrix $[\mathbf{n}^+]$ is defined as

$$[\mathbf{n}^+] = \begin{bmatrix} 0 & -n_3 & n_2 \\ n_3 & 0 & -n_1 \\ -n_2 & n_1 & 0 \end{bmatrix} , \quad (27.13)$$

and counterclockwise rotations are described by a positive angle ϕ . Expression (27.12) is used to describe the rotation in terms of four parameters, and it is known as *Euler axis-angle* attitude representation.

A closely related representation is the *Euler angles* parameterization. This follows from performing three consecutive rotations about the main axes of a frame, for which the rotation axis takes one of the forms $\mathbf{n}_1 = (1, 0, 0)^\top$, or $\mathbf{n}_2 = (0, 1, 0)^\top$,

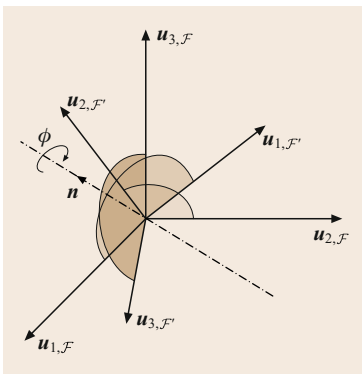


Fig. 27.7 Euler's theorem: in the three-dimensional space, any rotation of a rigid body can be described by a rotation of angle ϕ about an axis \mathbf{n}

or $\mathbf{n}_3 = (0, 0, 1)^\top$. Any arbitrary rotation in the three-dimensional space can be obtained by a sequence $i \rightarrow j \rightarrow k$, with $i \neq j$ and $j \neq k$. An example is the sequence 321, in which the first rotation is about axis $u_{3,\mathcal{F}}$ with angle ψ , the second is about the (new) axis $u_{2,\mathcal{F}'}$ with angle θ , and the last about the (new) axis $u_{1,\mathcal{F}''}$ with angle ϕ . Sequence 321, illustrated in Fig. 27.8, is of common use in vehicle guidance applications, where the angles ψ, θ, ϕ are named yaw, pitch and roll angles, respectively (an alternative definition of the triad is heading, elevation and bank). For such a combination, the attitude matrix reads

$$\mathbf{R}(\psi, \theta, \phi) = \mathbf{R}(u_1, \phi) \mathbf{R}(u_2, \theta) \mathbf{R}(u_3, \psi) = \begin{bmatrix} C_\psi C_\theta & +S_\psi C_\theta & -S_\theta \\ C_\psi S_\theta S_\phi - S_\psi C_\phi & S_\psi S_\theta S_\phi + C_\psi C_\phi & C_\theta S_\phi \\ C_\psi S_\theta C_\phi + S_\psi S_\phi & S_\psi S_\theta C_\phi - C_\psi S_\phi & C_\theta C_\phi \end{bmatrix} . \quad (27.14)$$

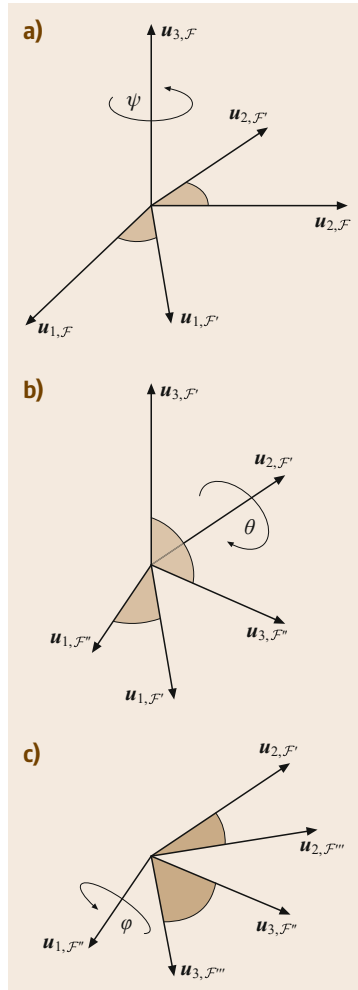


Fig. 27.8a-c Euler's angles sequence 321. The three consecutive rotation angles are commonly named yaw, pitch and roll. (a) First rotation with angle ψ about axis $u_{3,\mathcal{F}}$. (b) Second rotation with angle θ about axis $u_{2,\mathcal{F}'}$. (c) Third rotation with angle ϕ about axis $u_{1,\mathcal{F}''}$

The sequence of three rotations that form the attitude matrix \mathbf{R} is paramount, with different ordering yielding different attitude matrices.

Multiple Euler angles triads may represent the same orientation: this is usually avoided by limiting the parameter space. As an example, the sequence 321 is equivalently expressed as $\mathbf{R}(\psi, \theta, \varphi)$ and $\mathbf{R}(\psi + \pi, \pi - \theta, \varphi + \pi)$. Imposing $-\pi/2 < \theta \leq \pi/2$ and $-\pi < \psi, \varphi \leq \pi$ solves this multiplicity problem.

The Euler angles, although easily understood and thus of common use in attitude gages and indicators, are computationally less efficient and precise than other parameterizations, mainly due to the presence of trigonometric functions.

An alternative, widely used attitude parameterization is based on quaternion algebra [27.12, 13]. A quaternion is a real-valued, four-component entity \mathbf{q} , namely composed by a scalar part q_0 and a vectorial part $\mathbf{q} = (q_1, q_2, q_3)^\top$, that satisfies the following multiplication rule

$$\mathbf{q}\mathbf{q}' = q_0 \begin{pmatrix} q'_0 \\ \mathbf{q}' \end{pmatrix} + \begin{bmatrix} 0 & -\mathbf{q}^\top \\ \mathbf{q} & [\mathbf{q}^+] \end{bmatrix} \begin{pmatrix} q'_0 \\ \mathbf{q}' \end{pmatrix}, \quad (27.15)$$

where the skew-symmetric matrix $[\mathbf{q}^+]$ is obtained from the elements of the vectorial part of the quaternion as

$$[\mathbf{q}^+] = \begin{bmatrix} 0 & -q_3 & q_2 \\ q_3 & 0 & -q_1 \\ -q_2 & q_1 & 0 \end{bmatrix}. \quad (27.16)$$

Any quaternion of unit norm ($q_0^2 + \mathbf{q}^\top \mathbf{q} = 1$) represents a rotation, and it is defined as a *quaternion of rotation*. By linking the elements of a quaternion \mathbf{q} to the rotation axis and angle as

$$\mathbf{q}(\phi, \mathbf{n}) = \begin{pmatrix} \cos\left(\frac{\phi}{2}\right) \\ \sin\left(\frac{\phi}{2}\right) \mathbf{n} \end{pmatrix}, \quad (27.17)$$

the Euler rotation formula (27.12) becomes

$$\begin{aligned} \mathbf{R}(\mathbf{q}) &= (q_0^2 - \mathbf{q}^\top \mathbf{q}) \mathbf{I}_3 + 2\mathbf{q}\mathbf{q}^\top - 2q_0 [\mathbf{q}^+] = \\ &= \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 + q_0q_3) & 2(q_1q_3 - q_0q_2) \\ 2(q_1q_2 - q_0q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 + q_0q_1) \\ 2(q_1q_3 + q_0q_2) & 2(q_2q_3 - q_0q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix}. \end{aligned} \quad (27.18)$$

This is the *quaternion representation* of the attitude matrix. Note that $\mathbf{R}(\mathbf{q}) = \mathbf{R}(-\mathbf{q})$, thus any rotation can be

represented by two quaternions. The quaternion representation is computationally more efficient than other representations that employ trigonometric functions. The main disadvantages of using unit quaternions are the additional unit norm constraint and the lack of a direct physical interpretation. The latter is the reason why the attitude is usually output in human-machine interfaces in terms of Euler angles, even though the underlying control operations may be performed with quaternions.

Aiming to reduce the number of attitude parameters used in the quaternion representation, two alternative parameterizations can be used. The first is the *Rodrigues vector representation* (also known as Gibbs vector), which relates to the quaternion and Euler axis-angle representations as

$$\mathbf{p} = \frac{\mathbf{q}}{q_0} = \mathbf{n} \tan\left(\frac{\phi}{2}\right). \quad (27.19)$$

By definition, a rotation angle $\phi = \pi$ about any axis cannot be represented (singularity). The Rodrigues vector representation introduces some computational advantages, for example the inverse of $\mathbf{R}(\mathbf{p})$ is simply obtained as $\mathbf{R}(-\mathbf{p})$.

A second choice is given by the *modified Rodrigues representation*, alternatively defined as

$$\mathbf{p}^+ = \frac{\mathbf{q}}{1 + q_0}, \quad (27.20)$$

or

$$\mathbf{p}^- = \frac{\mathbf{q}}{1 - q_0}. \quad (27.21)$$

This modification of the Rodrigues vector is used to shift the singularity of the Rodrigues representation to 2π .

Finally, two additional attitude parameterizations can be used, both based on skew-symmetric matrices. The first is the *Cayley representation*, formulated from the skew-symmetric matrix associated to the Rodrigues vector \mathbf{p} as

$$\mathbf{R}([\mathbf{p}^+]) = (\mathbf{I}_3 - [\mathbf{p}^+])^{-1} (\mathbf{I}_3 + [\mathbf{p}^+]). \quad (27.22)$$

The second is the *skew representation*, formulated as a matrix exponentiation

$$\mathbf{R}([\mathbf{v}^+]) = \exp([\mathbf{v}^+]), \quad (27.23)$$

with \mathbf{v} any real-valued, three-component vector. The Euler rotation formula is obtained from (27.23) by replacing $\mathbf{v} = \mathbf{n}\phi$.

Each of attitude representations listed above provides the same attitude matrix \mathbf{R} for the same body orientation. However, each attitude parameterization differs in terms of computational load, singularities (orientations that cannot be parameterized), multiple

representations of the same attitude, and attitude estimation error in the parameter space. Further details about the attitude representations and their properties can be found in [27.14].

27.3 Attitude Estimation from Baseline Observations

The transformation between the coordinates of a vector expressed in two reference frames \mathcal{F} and \mathcal{F}' with common origin is realized by a rotation: $\mathbf{b}_{\mathcal{F}'} = \mathbf{R}\mathbf{b}_{\mathcal{F}}$. The solution of the inverse problem, that is, the determination of the relative orientation between frames \mathcal{F}' and \mathcal{F} from a set of baseline observations, is the objective of attitude estimation.

Before addressing the attitude estimation problem, a few definitions are given:

1. The *local* frame \mathcal{F} is the frame attached to the body whose attitude has to be determined. The *reference* frame \mathcal{B} is the frame used as a reference to compute the rotations.
2. The local baseline coordinates are cast in matrix

$$\mathbf{F} = [\mathbf{f}_1 \quad \mathbf{f}_2 \quad \cdots \quad \mathbf{f}_m]. \quad (27.24)$$

Each baseline \mathbf{f}_i is expressed in the local coordinate frame \mathcal{F} . Vector \mathbf{f}_i is a p -component vector, where the parameter p assumes one of the following values: $p = 1$ for any configuration of m aligned baselines; $p = 2$ for any noncollinear configurations of m coplanar baselines; $p = 3$ for any noncoplanar configurations of m baselines. The parameter p defines the rank of matrix \mathbf{F} : $p = \text{rk}(\mathbf{F})$. This definition of matrix \mathbf{F} enables solving the single-baselines attitude determination problem (Compass) within the same framework employed for arrays composed by a larger number of antennas.

3. The baseline coordinates in the reference frame \mathcal{B} are cast in matrix

$$\mathbf{B} = [\mathbf{b}_1 \quad \mathbf{b}_2 \quad \cdots \quad \mathbf{b}_m]. \quad (27.25)$$

Each baseline \mathbf{b}_i is a three-component vector.

4. The matrix of rotations \mathbf{R} has dimensions $3 \times p$, and is subject to the orthonormality constraint $\mathbf{R}^\top \mathbf{R} = \mathbf{I}_p$. In the following, $\mathcal{SO}(3, p) \subset \mathbb{R}^{3 \times p}$ denotes the group of *proper* $3 \times p$ rotation matrices, that is, orthonormal matrices with positive determinant. The positiveness of the matrix determinant is verified for $p = 2$ by adding the third (fully dependent) column of \mathbf{R} as $\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2$, whereas for $p = 1$ the constraint

is implicitly fulfilled. This generalization is introduced to address attitude estimation problems with single- or two-baseline observations [27.15].

27.3.1 Estimation of the Orthonormal Matrix of Rotations

Pseudorange and carrier-phase observations collected by the elements of a GNSS antenna array are processed to obtain baseline estimates, usually expressed in the Earth-Centered, Earth-Fixed (ECEF) frame or in the East-North-Up (ENU) frame. The link between the estimate of the baseline coordinates $\hat{\mathbf{B}}$ in these reference frames and the baseline coordinates \mathbf{F} in the local body-attached frame \mathcal{F} is

$$\hat{\mathbf{B}} = \mathbf{R}\mathbf{F} + \boldsymbol{\Theta} \quad ; \quad \mathbf{R} \in \mathcal{SO}(3, p). \quad (27.26)$$

Matrix $\boldsymbol{\Theta}$ denotes the baseline estimation error, and the unknown in (27.26) is the orthonormal matrix of rotations \mathbf{R} .

The simplest and oldest method to derive the attitude matrix is the three-axis attitude determination (TRIAD) method [27.16], which applies to two nonparallel baselines. The two baseline estimates and the local baseline vectors are cast in matrices

$$\hat{\mathbf{B}} = [\hat{\mathbf{b}}_1 \quad \hat{\mathbf{b}}_2]$$

and

$$\mathbf{F} = [\mathbf{f}_1 \quad \mathbf{f}_2].$$

Two orthonormal bases are then formed, $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ and $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$, with

$$\begin{aligned} \mathbf{v}_1 &= \frac{\hat{\mathbf{b}}_1}{\|\hat{\mathbf{b}}_1\|}, & \mathbf{u}_1 &= \frac{\mathbf{f}_1}{\|\mathbf{f}_1\|}, \\ \mathbf{v}_2 &= \frac{\hat{\mathbf{b}}_1 \times \hat{\mathbf{b}}_2}{\|\hat{\mathbf{b}}_1 \times \hat{\mathbf{b}}_2\|}, & \mathbf{u}_2 &= \frac{\mathbf{f}_1 \times \mathbf{f}_2}{\|\mathbf{f}_1 \times \mathbf{f}_2\|}, \\ \mathbf{v}_3 &= \mathbf{v}_1 \times \mathbf{v}_2, & \mathbf{u}_3 &= \mathbf{u}_1 \times \mathbf{u}_2. \end{aligned} \quad (27.27)$$

The sought orthonormal matrix is then completely defined by the nine direction cosines between the vectors

of the bases \mathcal{U} and \mathcal{V}

$$\hat{\mathbf{R}}_{\text{TRIAD}} = \sum_{i=1}^3 v_i \mathbf{u}_i^\top. \quad (27.28)$$

Although simple to implement, the TRIAD method only applies to baseline couples, and does not incorporate any weighting of the elements of the baseline estimates.

The problem of estimating the attitude matrix \mathbf{R} from an arbitrary set of baseline estimations $\hat{\mathbf{B}}$ can be formulated within the least-squares framework. The matrix of rotations is the orthonormal matrix that minimizes the squared weighted norm of the estimation errors $\Xi = \hat{\mathbf{B}} - \mathbf{R}\mathbf{F}$

$$\hat{\mathbf{R}} = \arg \min_{\mathbf{R} \in SO(3,p)} \|\text{vec}(\Xi)\|_{\mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}}^2, \quad (27.29)$$

where

$$\|\mathbf{x}\|_{\mathbf{Q}}^2 = \mathbf{x}^\top \mathbf{Q}^{-1} \mathbf{x}$$

denotes the squared weighted norm, vec is the operator that stacks the columns of a matrix $m \times n$, yielding the corresponding mn -vector, and $\mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}$ is the variance-covariance (v-c) matrix whose elements are the variances and covariances of and between the elements of the baseline estimation $\hat{\mathbf{B}}$. Expression (27.29) defines a nonlinear least-squares problem, where the nonlinearity is implicit to the orthonormality constraint. Depending on the shape of the weight matrix, different approaches for the minimization of (27.29) are available.

27.3.2 Orthogonal Procrustes Problem

If matrix $\mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}^{-1}$ is a block-diagonal matrix obtained as

$$\mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}^{-1} = \mathbf{W} \otimes \mathbf{I}_3,$$

with \mathbf{W} a diagonal $m \times m$ matrix and \otimes denoting the Kronecker product [27.17], expression (27.29) reduces to

$$\hat{\mathbf{R}} = \arg \min_{\mathbf{R} \in SO(3,p)} \text{tr} \left[\mathbf{W}^{\frac{1}{2}} \Xi^\top \Xi \mathbf{W}^{\frac{1}{2}} \right], \quad (27.30)$$

where tr denotes the trace operator. Problem (27.30) corresponds to those cases in which each baseline observation is weighted by its own factor w_i , and it is commonly referred to as Wahba's problem [27.18, 19], or the orthogonal Procrustes problem (OPP). The latter naming derives from the bandit Procrustes, who,

according to Greek mythology, used to capture passers-by, lie them on an iron bed, and forced them to fit the bed length by stretching or cutting off their limbs. The modern terminology *Procrustes analysis* refers to those problems in which Euclidean transformations, such as translation, rotation and scaling, are applied to an object in order to fit predetermined constraints.

The solution of (27.26) is found by maximizing the following expression

$$\hat{\mathbf{R}} = \arg \max_{\mathbf{R} \in SO(3,p)} \text{tr} \left[\mathbf{R} \mathbf{F} \mathbf{W} \hat{\mathbf{B}}^\top \right]. \quad (27.31)$$

If the matrix product $\mathbf{F} \mathbf{W} \hat{\mathbf{B}}^\top$ is nonsingular, its singular value decomposition (SVD)

$$\mathbf{F} \mathbf{W} \hat{\mathbf{B}}^\top = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$$

enables rewriting the term to maximize as

$$\sum_{i=1}^n x_{ii} d_i,$$

with x_{ii} the diagonal elements of matrix $\mathbf{X} = \mathbf{V}^\top \mathbf{R} \mathbf{U}$. Since \mathbf{X} is a product of orthonormal matrices, and the determinant of $\hat{\mathbf{R}}$ must be positive, the maximizer of (27.31) is

$$\hat{\mathbf{R}} = \mathbf{V} \begin{bmatrix} \det(\mathbf{V}^\top \mathbf{U}) & 0 \\ 0 & \mathbf{I}_2 \end{bmatrix} \mathbf{U}^\top. \quad (27.32)$$

Expression (27.32) solves the attitude estimation problem (27.30) with the computation of a SVD [27.18]. The latter can be avoided by following an elegant reasoning based on the Davenport's *q-method* [27.20], which redefines the maximization problem (27.31) in terms of the quaternion used to represent the attitude matrix

$$\hat{\mathbf{q}} = \arg \max_{\|\mathbf{q}\|=1} (\mathbf{q}_0, \mathbf{q}^\top)^\top \mathbf{K} (\mathbf{q}_0, \mathbf{q}^\top)^\top, \quad (27.33)$$

where the *Davenport matrix* \mathbf{K} is built as

$$\begin{aligned} \mathbf{K} &= \begin{bmatrix} \mathbf{C} + \mathbf{C}^\top - \text{tr}(\mathbf{C})\mathbf{I}_3 & \mathbf{s} \\ \mathbf{s}^\top & \text{tr}(\mathbf{C}) \end{bmatrix}, \\ \mathbf{C} &= \hat{\mathbf{B}} \mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}^{-1} \mathbf{F}^\top, \\ \mathbf{s} &= (c_{2,3} - c_{3,2}, \quad c_{3,1} - c_{1,3}, \quad c_{1,2} - c_{2,1})^\top, \end{aligned} \quad (27.34)$$

with $c_{i,j}$ denoting the i, j component of matrix \mathbf{C} . The unknown is then the quaternion that maximizes the

bilinear term in (27.33), subject to the unit norm constraint, and it is found by selecting the eigenvector associated to the largest eigenvalue ξ_{\max} of \mathbf{K} . The computational load for the eigenvalue decomposition of matrix \mathbf{K} is comparable to the one of the SVD [27.21], but a very fast algorithm can nevertheless be devised, which avoids the direct calculation of the eigenvalues of \mathbf{K} . When precise baseline estimates are available, the largest eigenvalue ξ_{\max} is close to the approximate solution

$$\xi_{\max} \approx \xi_0 = \frac{1}{2} \text{tr} \left[\mathbf{W}^{\frac{1}{2}} \left[\hat{\mathbf{B}}^{\top} \hat{\mathbf{B}} + \mathbf{F}^{\top} \mathbf{F} \right] \mathbf{W}^{\frac{1}{2}} \right]. \quad (27.35)$$

The secular – or characteristic – fourth-order equation

$$f(\xi) = \det(\mathbf{K} - \xi \mathbf{I}_4) = 0, \quad (27.36)$$

with ξ_0 as the initialization point, is then solved to extract the largest root, thus avoiding the computation of any decomposition [27.22].

Several methods exploiting the quaternion parameterization approach have been formulated and widely used in practice, such as the quaternion estimator (QUEST) [27.22, 23], the fast optimal attitude matrix (FOAM) [27.24], the estimator of the optimal quaternion (ESOQ) [27.25], and the second ESOQ (ESOQ2) [27.26] algorithms. These approaches, which only differ in the way the characteristic equation (27.36) is handled, are extremely numerically efficient, and as such are suitable for implementation also in low-grade processors [27.27–29].

27.3.3 Weighted Orthogonal Procrustes Problem

If matrix $\mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}^{-1}$ can be decomposed as

$$\mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}^{-1} = \mathbf{I}_m \otimes \mathbf{\Gamma},$$

with $\mathbf{\Gamma}$ denoting a diagonal 3×3 matrix, expression (27.29) reduces to

$$\hat{\mathbf{R}} = \arg \min_{\mathbf{R} \in \mathcal{SO}(3,p)} \text{tr} [\mathbf{\Xi}^{\top} \mathbf{\Gamma} \mathbf{\Xi}]. \quad (27.37)$$

In the minimization problem (27.37), each row of the baseline matrix $\hat{\mathbf{B}}$ is weighted differently. This is a viable weighting approach when operating with sensors that provide higher accuracies in certain directions, but does not give proper weighting to all the elements of the baseline estimates. Expression (27.37) defines a weighted orthogonal Procrustes problem (WOPP), and although less complex than the case of a fully populated matrix $\mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}$, it can only be solved with numerical techniques. Efficient numerical schemes to guarantee fast convergence to a global minimum are reviewed in [27.30, 31].

27.3.4 Attitude Estimation with Fully Populated Weight Matrix

Generally, matrix $\mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}^{-1}$ cannot be reduced to one of the previous cases when working with GNSS baseline estimates. First, it is useful to operate a quadratic decomposition of the squared weighted norm in (27.29) as [27.32]

$$\begin{aligned} \left\| \text{vec}(\hat{\mathbf{B}} - \mathbf{R}\mathbf{F}) \right\|_{\mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}}^2 \\ = \left\| \text{vec}(\hat{\mathbf{E}}) \right\|_{\mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}}^2 + \left\| \text{vec}(\hat{\mathbf{R}} - \mathbf{R}) \right\|_{\mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}}}^2, \end{aligned} \quad (27.38)$$

where $\hat{\mathbf{E}} = \hat{\mathbf{B}} - \hat{\mathbf{R}}\mathbf{F}$, and

$$\begin{aligned} \hat{\mathbf{R}} &= \mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}}(\mathbf{F}^{\top} \otimes \mathbf{I}_3)^{\top} \mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}^{-1} \text{vec}(\hat{\mathbf{B}}), \\ \mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}} &= \left[(\mathbf{F}^{\top} \otimes \mathbf{I}_3)^{\top} \mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}^{-1} (\mathbf{F}^{\top} \otimes \mathbf{I}_3) \right]^{-1}. \end{aligned} \quad (27.39)$$

Matrices $\hat{\mathbf{R}}$ and $\mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}}$ denote the least-squares estimation of the attitude matrix and the corresponding v-c matrix when the orthonormality constraint is discarded in the estimation problem (27.29). The orthonormal matrix of rotations is then found by minimizing the squared weighted norm

$$\tilde{\mathbf{R}} = \arg \min_{\mathbf{R} \in \mathcal{SO}(3,p)} \left\| \text{vec}(\hat{\mathbf{R}} - \mathbf{R}) \right\|_{\mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}}}^2, \quad (27.40)$$

where the variances of the elements of $\hat{\mathbf{R}}$ and their correlation are fully accounted for when weighting the attitude estimation residuals $\hat{\mathbf{R}} - \mathbf{R}$.

The sought orthonormal matrix is obtained by projecting the vector $\text{vec}(\hat{\mathbf{R}})$, element of the $3p$ -dimensional space, onto the curved manifold in the $0.5p(p+1)$ -dimensional subspace defined by the constraints $\mathbf{R}^{\top} \mathbf{R} = \mathbf{I}_p$, where the metric of the projection is defined by matrix $\mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}}$. Figure 27.9 visualizes a single-baseline example ($p = 1$), where vector $\hat{\mathbf{r}}$ is projected onto the sphere $\mathbb{S} := \mathbf{r}^{\top} \mathbf{r} = 1$. The sought attitude solution $\tilde{\mathbf{r}}$ is the point of contact between the sphere \mathbb{S} and the ellipsoid \mathbb{E} centered in $\hat{\mathbf{r}}$ and shaped by the elements of matrix $\mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}}$.

The solution of the nonlinear constrained problem (27.40) can be computed via the Lagrangian multipliers method [27.33]. Alternative numerical solutions can be devised by employing a parameterization of the attitude matrix. For example, by representing matrix \mathbf{R} with a triplet of Euler angles, the orthonormality constraints are implicitly fulfilled, and expression (27.40) reduces to a nonlinear least-squares minimization problem that can be solved via the Newton method or, disregarding the second-order derivatives, the Gauss–Newton method [27.34].

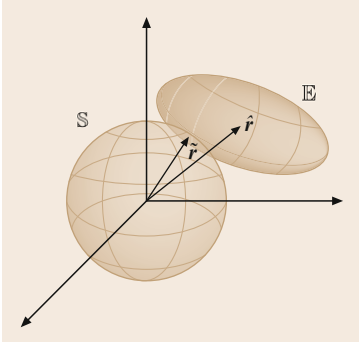


Fig. 27.9 The point of contact between the unit sphere \mathbb{S} and the ellipsoid \mathbb{E} , centered in $\hat{\mathbf{r}}$ and shaped by $\mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}}$, is the solution of the constrained least-squares problem (27.40) in the single-baseline case

27.3.5 On the Precision of Attitude Estimation

The precision of angular estimations from baseline observations depends on two factors: the length of the baselines employed and the baseline observation noise. Assuming that precise baseline observations are available, such that the dispersion of $\hat{\mathbf{R}}$ in (27.39) is small, a first-order description of the attitude estimation error is given by the v-c matrix $\mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}}$, whose entries depend on the quality of the baseline estimates through matrix $\mathbf{Q}_{\hat{\mathbf{B}}\hat{\mathbf{B}}}$ and on the geometrical properties of the antenna array through matrix \mathbf{F} . Relation (27.39) shows how these factors affect the quality of the GNSS attitude estimates. The attitude estimation error is inversely proportional to the geometrical size of the array (baseline lengths),

and directly proportional to the observation noise. The latter can be effectively improved by exploiting GNSS carrier-phase measurements, as shown in more detail in Sect. 27.4.4.

Due to the nonlinear constraints in (27.40), it is not possible to obtain a measure of the attitude estimation error by linearly propagating the float v-c matrix $\mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}}$. The dispersion of the orthonormal matrix of rotations \mathbf{R} can no longer be described by a normal distribution. An approximated measure of the estimation error can be computed by linearizing the function that maps matrix $\hat{\mathbf{R}}$ into the matrix of rotations $\tilde{\mathbf{R}}$ [27.22].

In many applications, an indication of the estimation error in the attitude parameters domain is preferred over knowing the precision of the elements of matrix $\hat{\mathbf{R}}$. Let $\boldsymbol{\gamma}$ denote the c -component vector of attitude parameters chosen to represent the matrix of rotations (e.g., $c = 3$ for the Euler angles and $c = 4$ for the quaternion of rotation). The attitude parameters are in nonlinear correspondence with the elements of the attitude matrix, and their precision can be described with a first-order approximation by the v-c matrix $\mathbf{Q}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$, obtained as

$$\mathbf{Q}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = \mathbf{J}_{\boldsymbol{\gamma}(\mathbf{R})} \mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}} \mathbf{J}_{\boldsymbol{\gamma}(\mathbf{R})}^{\top}, \quad (27.41)$$

with $\mathbf{J}_{\boldsymbol{\gamma}(\mathbf{R})}$ denoting the $c \times 3p$ Jacobian matrix of the inverse function that maps from the elements of the attitude matrix to the chosen space of attitude parameters.

27.4 The GNSS Attitude Model

The previous two sections reviewed how the orientation of a frame can be described via a set of attitude parameters and how these are estimated from baseline measurements. In the context of GNSS-based attitude sensors, the (vectorial) baselines between the elements of the antenna array must be estimated from the pseudorange and carrier-phase measurements. A coherent theoretical framework for GNSS attitude determination is built by formulating a functional model that links the GNSS observables with the sought attitude parameters, while embedding all the available a priori geometrical constraints.

First, differential pseudorange and carrier-phase measurements are formed between two closely-separated (i.e., with a baseline not exceeding a few hundred meters) antennas tracking the same satellite, such that the differential ionospheric and tropospheric delays can be neglected (Chap. 20). The single difference (SD) measurements are free from the (common) satellite clock error, but may still contain differential

receiver-clock errors and instrumental delays. A synchronization between the clocks at the two receivers and a correct calibration of the line biases is necessary to augment the strength of the observation model, without the necessity of introducing additional unknowns. An alternative method to eliminate the remaining differential biases is to form double difference (DD) measurements, in which the only remaining unknowns are the integer ambiguities and the baseline coordinates.

Assuming normally distributed measurements collected at n channels (SD or DD) on f frequencies, the following single-baseline observation model holds for any GNSS used

$$\mathbf{y} \sim \mathcal{N}(\mathbf{A}\mathbf{z} + \mathbf{G}\mathbf{b}, \mathbf{Q}_{\mathbf{y}\mathbf{y}}), \quad \mathbf{z} \in \mathbb{Z}^{nf}, \quad \mathbf{b} \in \mathbb{R}^3, \quad (27.42)$$

with \mathbf{y} the $2nf$ -component differential data vector; \mathbf{A} the $2nf \times 2nf$ design matrix defined as $\mathbf{A} = (0, 1)^{\top} \otimes \mathbf{\Lambda} \otimes \mathbf{I}_n$, with $\mathbf{\Lambda}$ the diagonal $f \times f$ matrix containing the wavelength(s); \mathbf{G} the $2nf \times 3$ design matrix constructed

as $\mathbf{G} = \mathbf{e}_{2f} \otimes \mathbf{U}$, where \mathbf{e}_{2f} denotes a $2f$ -component vector of ones and \mathbf{U} is the matrix of differenced line-of-sight unit vectors. The unknown parameters are the integer-valued vector of nf carrier-phase ambiguities \mathbf{z} and the real-valued 3-component vector of baseline coordinates \mathbf{b} . To ease the notation, the time dependency is not explicitly indicated.

The data vector \mathbf{y} is generally assumed to be normally distributed, with dispersion captured by the $2nf \times 2nf$ v-c matrix \mathbf{Q}_{yy} .

The linear model (27.42) is generalized for an array of $m+1$ GNSS antennas forming m baselines as

$$\begin{aligned} \text{vec}(\mathbf{Y}) &\sim \mathcal{N}(\text{vec}(\mathbf{AZ} + \mathbf{GB}), \mathbf{Q}_{YY}) \\ \mathbf{Z} &\in \mathbb{Z}^{nf \times m}, \mathbf{B} \in \mathbb{R}^{3 \times m}. \end{aligned} \quad (27.43)$$

The m vectors of $2nf$ differential observations of type (27.42) are the columns of matrix \mathbf{Y} ; the m carrier-phase integer ambiguity vectors are the columns of matrix \mathbf{Z} and the m baseline coordinates vectors are the columns of matrix \mathbf{B} . Matrix \mathbf{A} remains unchanged with respect to the single-baseline model (27.42). Also, it is assumed that the same geometry matrix \mathbf{G} applies to each baseline of the array. This is justified by considering a sufficiently small array size, with separations between antennas negligible with respect to the satellite-to-receivers distances (approximately 20 000 km).

The stochastic properties of the measurement vector $\text{vec}(\mathbf{Y})$ are described by the v-c matrix \mathbf{Q}_{YY} . If each of the m baseline measurements is fittingly described by the same v-c matrix \mathbf{Q}_{yy} , and the baselines are formed by using a common reference antenna, matrix \mathbf{Q}_{YY} can be reduced to

$$\mathbf{Q}_{YY} = \mathbf{P}_m \otimes \mathbf{Q}_{yy}, \quad (27.44)$$

where \mathbf{P}_m denotes the $m \times m$ matrix that introduces the correct covariance factors due to the differencing with respect to a common antenna. This matrix is computed as

$$\mathbf{P}_m = \frac{1}{2} (\mathbf{I}_m + \mathbf{e}_m \mathbf{e}_m^T),$$

with \mathbf{e}_m denoting an m -component vector of ones.

In a GNSS antenna array operating as an attitude sensor, the differential pseudorange and carrier-phase measurements are obtained from antennas at known relative positions. Under the rigid body hypothesis, with invariant relative positions between the elements of the array, the unknown baseline coordinates \mathbf{B} in (27.43) are obtained through a rotation of the known local baseline coordinates \mathbf{F} . The baseline coordinates in \mathbf{B} are expressed in a user-defined reference frame, such as the

ECEF or ENU frames, whereas the known local baseline coordinates \mathbf{F} are expressed in a reference frame attached to the body. The attitude of the body is estimated by computing the rotation matrix that describes the mutual orientation between the two frames. The GNSS attitude observation model is then formulated as [27.15]

$$\begin{aligned} \text{vec}(\mathbf{Y}) &\sim \mathcal{N}(\text{vec}(\mathbf{AZ} + \mathbf{GRF}), \mathbf{Q}_{YY}) \\ \mathbf{Z} &\in \mathbb{Z}^{nf \times m}, \mathbf{R} \in \mathcal{SO}(3, p), \end{aligned} \quad (27.45)$$

where the unknowns are the integer-valued matrix of carrier-phase ambiguities \mathbf{Z} and the rotation matrix \mathbf{R} .

The GNSS attitude model (27.45) shows several useful properties. First, the number of unknown parameters does not linearly increase with the number of baselines employed, with a gain in the observations-to-unknowns redundancy equal to $3(m-p)$. For example, five GNSS antennas forming four noncoplanar baselines give twelve unknown baseline coordinates, but only nine unknown entries of \mathbf{R} in (27.45). Moreover, due to the orthonormality constraint only three independent parameters of \mathbf{R} have to be estimated, further increasing the observations-to-unknowns redundancy.

Second, the actual number of constraints associated to the transformation $\mathbf{B} = \mathbf{RF}$ may be larger than the $\frac{1}{2}p(p+1)$ nonlinear constraints that follow from the orthonormality of \mathbf{R} [27.35]. The nonlinear constraints are explicit and always present, and their maximum number is six when employing four or more noncoplanar antennas. In addition, implicit $3(m-p)$ linear constraints exist when using three or more collinear antennas ($p=1, m \geq 2$), four or more coplanar antennas ($p=2, m \geq 3$), or any configuration of five or more antennas ($m > 3$). These implicit, linear constraints contribute to improve the overall estimation quality.

27.4.1 Potential Model Errors and Misspecification

The GNSS attitude model given in (27.45) neglects a number of parameters whose contributions may become significant, such as multipath, atmospheric delays, or antenna phase center variations.

Multipath accounts for the largest part of the measurement error in short baseline models. Multipath causes a degradation of the tracking capabilities of the receiver, which processes the sum of the direct signal and delayed signal replicas due to reflections and/or additional propagation paths. The multipath error depends on both magnitude and delay of the spurious contributions, which in turn depend on the relative geometry between antennas, GNSS satellites and surrounding objects, on the amplitude, frequency and polarization of the received signal, and on the platform dynamics.

Multipath does not behave as a random noise source, but rather as a time-changing observation bias, with dynamics proportional to the relative motion between satellite and receiving antenna. Multipath affects both code and phase tracking, introducing errors in pseudorange and carrier-phase measurements reaching meter-level and cm-level, respectively, which cannot be easily eliminated or estimated (Chap. 15). Techniques for the mitigation of multipath span from ad hoc antenna siting and physical design (e.g., choke rings, see Chap. 17) to narrow correlators, which attenuate the impact of reflected signals through narrower correlator spacing [27.36]. Also, the new generation of GNSS signals, based on the multiplexed binary offset carrier (MBOC) modulation, promises an improved multipath suppression [27.37, 38].

The atmospheric delays are usually neglected in differential applications employing short baselines, as the signals collected by antennas placed no further than a few hundred meters apart have traveled to a large extent the same path, thus experiencing similar delays that cancel out for the most part after differencing. However, the presence of strong ionospheric perturbations, such as scintillations or gradients, cannot be excluded a priori. These disturbances could alter the ranging signals introducing nonnegligible delays, loss of locks and additional observation noise.

An additional, potential source of estimation biases is the variation of the antenna phase centers (APCs). The signals collected at each antenna are by convention referenced to the corresponding APC (Chap. 17). The a priori knowledge of the distance between APCs is used to form the matrix of local baseline coordinates \mathbf{F} . Any modeling error of the local baseline coordinates, due to a wrong calibration of the APCs or due to their time-varying behavior not being accounted for, would directly translate into model misspecifications, possibly causing incorrect ambiguity resolution and attitude estimation biases. Variations of the APCs can also be caused by nonrigid bodies, in which the relative positions between the antennas change without the deformation being captured by the corresponding matrix of known local baseline coordinates.

A correct calibration and a suitable error detection and model misspecification testing procedures (Chap. 24) are usually implemented to avoid estimation biases due to the aforementioned error sources.

27.4.2 Resolution of the GNSS Attitude Model

The GNSS attitude model (27.45) links the measurement matrix to the unknown integer ambiguities and orthonormal rotation matrix. Although the relation-

ship between measurements and unknowns is linear, the associated constraints yield a nonlinear estimation problem. Precise GNSS-based attitude estimations can only be achieved by fixing the carrier-phase ambiguities to their correct integer values. When this is accomplished successfully, the differential carrier-phase observations act as very precise differential range measurements, making available accurate baseline estimations that are used to precisely estimate the body attitude. Integer ambiguity resolution in the presence of geometrical constraints is a complex problem, even more so when aiming at reliable and fast methods that could provide instantaneous solutions, without requiring long observation time spans or ad hoc initializations.

Several ambiguity resolution methods that apply to the GNSS-based attitude determination problem have been proposed. A first, straightforward method is to accumulate carrier-phase observations over time, during which the integer ambiguities remain constant, until the decoupling between ambiguities and attitude parameters enables their separate estimation [27.39]. This technique requires a certain amount of relative motion during the observation time span, since it relies on variations of the line-of-sight vectors between the antenna array and the GNSS satellites. Several methods that exploit this principle have subsequently been developed – and are commonly referred to as motion-based methods – aiming to limit the amount of relative motion needed for convergence [27.40], to eliminate the need for an a priori attitude estimate [27.41], or to improve the convergence time by also performing a search in the space of the unknown ambiguity-attitude parameters [27.42–45].

A common, potential drawback of all motion-based procedures is that the relative motion may take some time to occur. These techniques may then suffer from long initialization times, especially in low-dynamic applications, and cannot provide an instantaneous solution. This shortfall of motion-based approaches is resolved by methods that can provide an instantaneous solution of the unknown ambiguity and attitude parameters. These methods minimize a given cost function by performing an exhaustive search in the space of the unknown parameters. The search applies either in the position domain, over the space of relative antenna positions, or in the ambiguity domain, over the space of integer vectors.

Examples of approaches performing a search in the position domain include those based on the ambiguity function method (AFM), which minimizes a cost function parameterized in terms of direction angles [27.46]. The function is evaluated over a grid of potential azimuth-elevation angles in single-baseline

arrays [27.47–49] or over a grid of potential three-axis attitude angles for multiple-baseline arrays [27.50]. However, due to the dense quantization of the search domain necessary to find the global minimum, the computation of a very large number of samples of the objective function is required. The computational effort demanded by these approaches may thus hinder real-time applications [27.44].

The alternative methods of searching in the integer domain explore a number of potential integer ambiguity solutions by evaluating whether the corresponding baseline estimations satisfy the known geometrical constraints. Some approaches search for a candidate integer vector [27.51–54], or a selected subset of the integer vector [27.55–57], and discard those candidates that do not generate baseline solutions lying on a sphere of known radius. Other methods also weigh the attitude solution that follows from estimating multiple baselines based on different candidate integer vectors, for example [27.58].

Several integer search-based approaches make use of the least-squares ambiguity decorrelation adjustment (LAMBDA) method [27.59, 60], thus applying the integer least-squares (ILS) principle to minimize over the integer domain (Chap. 23). The method is numerically efficient [27.61, 62], and it is demonstrated to be optimal, in terms of providing the highest probability of correct ambiguity resolution when applied to linear models [27.63, 64]. The LAMBDA method is widely used in the context of instantaneous GNSS-based attitude determination applications, see for example [27.51–54, 65, 66]. Many methods exploit the geometrical constraints by rejecting unlikely baseline solutions. This use of a priori known information is legitimate, and yields enhanced success rates, but does not guarantee to minimize the least-squares residual error. The optimal solution to the problem of joint ambiguity and attitude estimation is derived from the rigorous inclusion of the geometrical constraints into the integer ambiguity resolution process. The resulting nonlinear, mixed estimation problem can be resolved by an extension of the ILS principle, implemented in the multivariate constrained-LAMBDA (MC-LAMBDA) method [27.15, 67–69], in which the nonlinear constraints are fully integrated into the cost function. The constraints play an active role during the integer search, providing guidance towards an instantaneous, joint ambiguity-attitude solution.

27.4.3 The GNSS Ambiguity and Attitude Estimation

The solution of the GNSS attitude model (27.45) is found by addressing the associated least-squares prob-

lem with inclusion of the integer constraint on the carrier-phase ambiguities \mathbf{Z} and of the orthonormality constraint on the attitude matrix \mathbf{R}

$$\{\check{\mathbf{Z}}, \check{\mathbf{R}}\} = \arg \min_{\substack{\mathbf{Z} \in \mathbb{Z}^{nf \times m}, \\ \mathbf{R} \in SO(3,p)}} \|\text{vec}(\mathbf{Y} - \mathbf{AZ} - \mathbf{GRF})\|_{\mathbf{Q}_{\mathbf{Y}\mathbf{Y}}}^2. \quad (27.46)$$

This solution is optimal in a least-squares sense, that is, the squared weighted norm of the estimation residuals is minimized over the parameter space. The minimization problem (27.46) is resolved with a two-steps procedure. First, a so-called *float solution* is derived by disregarding both the integer and the orthonormality constraints. The precision of this unconstrained estimation is driven by the precision of the pseudorange measurements, thereby acting as an initial coarse solution. The float estimation is refined in a second step by jointly estimating the integer ambiguities and the matrix of rotations, accounting for both the integer and orthonormality constraints. The two steps of the resolution procedure are detailed as follows.

Step I: Float Solution

The float solution of model (27.45) is the solution of the least-squares minimization problem

$$\{\hat{\mathbf{Z}}, \hat{\mathbf{R}}\} = \arg \min_{\substack{\mathbf{Z} \in \mathbb{R}^{nf \times m}, \\ \mathbf{R} \in \mathbb{R}^{3 \times p}}} \|\text{vec}(\mathbf{Y} - \mathbf{AZ} - \mathbf{GRF})\|_{\mathbf{Q}_{\mathbf{Y}\mathbf{Y}}}^2. \quad (27.47)$$

In this step, the estimation of the carrier-phase ambiguities is real-valued, and the matrix of rotations is not orthonormally constrained. The least-squares estimations $\hat{\mathbf{Z}}$ and $\hat{\mathbf{R}}$ are obtained by solving the associated set of normal equations

$$\mathbf{M} \begin{pmatrix} \text{vec}(\hat{\mathbf{Z}}) \\ \text{vec}(\hat{\mathbf{R}}) \end{pmatrix} = \begin{bmatrix} \mathbf{I}_m \otimes \mathbf{A}^\top \\ \mathbf{F} \otimes \mathbf{G}^\top \end{bmatrix} \mathbf{Q}_{\mathbf{Y}\mathbf{Y}}^{-1} \text{vec}(\mathbf{Y}), \quad (27.48)$$

with

$$\mathbf{M} = \begin{bmatrix} \mathbf{I}_m \otimes \mathbf{A}^\top \\ \mathbf{F} \otimes \mathbf{G}^\top \end{bmatrix} \mathbf{Q}_{\mathbf{Y}\mathbf{Y}}^{-1} [\mathbf{I}_m \otimes \mathbf{A} \quad \mathbf{F}^\top \otimes \mathbf{G}]. \quad (27.49)$$

For a measurement v-c matrix constructed as $\mathbf{Q}_{\mathbf{Y}\mathbf{Y}} = \mathbf{P}_m \otimes \mathbf{Q}_{yy}$, the float estimations $\hat{\mathbf{Z}}$ and $\hat{\mathbf{R}}$ are explicitly computed as

$$\begin{aligned} \hat{\mathbf{R}} &= [\bar{\mathbf{G}}^\top \mathbf{Q}_{yy}^{-1} \bar{\mathbf{G}}]^{-1} \bar{\mathbf{G}}^\top \mathbf{Q}_{yy}^{-1} \mathbf{Y} \mathbf{P}_m^{-1} \mathbf{F}^\top [\mathbf{F} \mathbf{P}_m^{-1} \mathbf{F}^\top]^{-1} \\ \hat{\mathbf{Z}} &= [\mathbf{A}^\top \mathbf{Q}_{yy}^{-1} \mathbf{A}]^{-1} \mathbf{A}^\top \mathbf{Q}_{yy}^{-1} [\mathbf{Y} - \mathbf{G} \hat{\mathbf{R}} \mathbf{F}], \end{aligned} \quad (27.50)$$

with $\tilde{\mathbf{G}} = \mathbf{P}_A^\perp \mathbf{G}$. Matrix $\mathbf{P}_C^\perp = \mathbf{I} - \mathbf{P}_C$ denotes the orthogonal complement of the projector matrix

$$\mathbf{P}_C = \mathbf{C} [\mathbf{C}^\top \mathbf{Q}_{yy}^{-1} \mathbf{C}]^{-1} \mathbf{C}^\top \mathbf{Q}_{yy}^{-1}.$$

The float estimations are normally distributed, with dispersion computed by inverting the normal matrix \mathbf{M}

$$\begin{bmatrix} \mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{Z}}} & \mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{R}}} \\ \mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{Z}}} & \mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}} \end{bmatrix} = \mathbf{M}^{-1}. \quad (27.51)$$

Step II: Ambiguity-Attitude Estimation

In this second step the float solution is refined by estimating the integer carrier-phase ambiguities. The squared norm in (27.46) is decomposed in terms of a sum-of-squares as [27.32]

$$\begin{aligned} & \|\text{vec}(\mathbf{Y} - \mathbf{AZ} - \mathbf{GRF})\|_{\mathbf{Q}_{yy}}^2 \\ &= \|\text{vec}(\boldsymbol{\Psi})\|_{\mathbf{Q}_{yy}}^2 + \|\text{vec}(\hat{\mathbf{Z}} - \mathbf{Z})\|_{\mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{Z}}}}^2 \\ &+ \|\text{vec}(\hat{\mathbf{R}}(\mathbf{Z}) - \mathbf{R})\|_{\mathbf{Q}_{\hat{\mathbf{R}}(\mathbf{Z})\hat{\mathbf{R}}(\mathbf{Z})}}^2, \end{aligned} \quad (27.52)$$

where matrix $\boldsymbol{\Psi}$ denotes the float estimation error, matrix $\hat{\mathbf{R}}(\mathbf{Z})$ denotes the conditional attitude matrix, that is, the (generally not orthonormal) attitude matrix conditioned on the ambiguity residual $\text{vec}(\hat{\mathbf{Z}} - \mathbf{Z})$, and matrix $\mathbf{Q}_{\hat{\mathbf{R}}(\mathbf{Z})\hat{\mathbf{R}}(\mathbf{Z})}$ denotes the v-c matrix of the conditional solution. The conditional attitude matrix $\hat{\mathbf{R}}(\mathbf{Z})$ is computed by considering the integer ambiguities as known and solving the associated (unconstrained) model $\mathbf{E}(\mathbf{Y} - \mathbf{AZ}) = \mathbf{GRF}$ in a least-squares sense, obtaining

$$\begin{aligned} \text{vec}(\hat{\mathbf{R}}(\mathbf{Z})) &= \text{vec}(\hat{\mathbf{R}}) - \mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{Z}}} \mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{Z}}}^{-1} \text{vec}(\hat{\mathbf{Z}} - \mathbf{Z}), \\ \mathbf{Q}_{\hat{\mathbf{R}}(\mathbf{Z})\hat{\mathbf{R}}(\mathbf{Z})} &= \mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}} - \mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{Z}}} \mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{Z}}}^{-1} \mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{R}}}. \end{aligned} \quad (27.53)$$

The conditional attitude solution is largely more precise than its float solution, as it mainly depends on the precision of the carrier-phase observables (Sect. 27.4.4).

The ambiguity-attitude estimation problem that follows from (27.52) reads [27.15]

$$\begin{aligned} \check{\mathbf{Z}} &= \arg \min_{\mathbf{Z} \in \mathbb{Z}^{nf \times m}} C(\mathbf{Z}), \\ C(\mathbf{Z}) &= \|\text{vec}(\hat{\mathbf{Z}} - \mathbf{Z})\|_{\mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{Z}}}}^2 + J(\mathbf{Z}). \end{aligned} \quad (27.54)$$

The function to be minimized is the sum of two dependent terms. The first term is the square distance between

the integer candidate \mathbf{Z} to the ambiguity float solution $\hat{\mathbf{Z}}$, in the metric of the v-c matrix $\mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{Z}}}$. The second term $J(\mathbf{Z})$ is computed as

$$\begin{aligned} J(\mathbf{Z}) &= \|\text{vec}(\hat{\mathbf{R}}(\mathbf{Z}) - \check{\mathbf{R}}(\mathbf{Z}))\|_{\mathbf{Q}_{\hat{\mathbf{R}}(\mathbf{Z})\hat{\mathbf{R}}(\mathbf{Z})}}^2, \\ \check{\mathbf{R}}(\mathbf{Z}) &= \arg \min_{\mathbf{R} \in SO(3 \times p)} \|\text{vec}(\hat{\mathbf{R}}(\mathbf{Z}) - \mathbf{R})\|_{\mathbf{Q}_{\hat{\mathbf{R}}(\mathbf{Z})\hat{\mathbf{R}}(\mathbf{Z})}}^2, \end{aligned} \quad (27.55)$$

and weighs the squared distance between the conditional float solution $\hat{\mathbf{R}}(\mathbf{Z})$ and the corresponding orthonormal solution $\check{\mathbf{R}}(\mathbf{Z})$, in the metric of $\mathbf{Q}_{\hat{\mathbf{R}}(\mathbf{Z})\hat{\mathbf{R}}(\mathbf{Z})}$.

The minimization of the cost function in (27.54) entails an integer search, in which the evaluation of $C(\mathbf{Z})$ at point \mathbf{Z} requires the corresponding attitude solution $\hat{\mathbf{R}}(\mathbf{Z})$ to be derived. This requires solving the nonlinear constrained problem in (27.55), whose solution was discussed in Sect. 27.3. Integer ambiguities that return unlikely attitude solutions – in terms of the weighted distance to the corresponding float estimation – yield proportionally larger values for the objective function $C(\mathbf{Z})$.

The integral ambiguity-attitude minimization problem (27.54) is solved by applying an extension of the ILS principle (Chap. 23), modified so as to deal with the additional nonlinear geometrical constraints, which largely strengthen the estimation procedure. The integer-valued minimizer $\check{\mathbf{Z}}$ is searched within a subset of the integer domain $\mathbb{Z}^{nf \times m}$: the search space is defined as

$$S(\chi^2) = \{\mathbf{Z} \in \mathbb{Z}^{nf \times m} \mid C(\mathbf{Z}) \leq \chi^2\}, \quad (27.56)$$

with χ^2 a positive scalar chosen to limit the branches of the search tree.

The integer search in the constrained ILS approach is numerically complex. This is due to a combination of two factors. First, the evaluation of the cost function $C(\mathbf{Z})$ requires the solution of the constrained minimization problem (27.55). Second, the scalar χ^2 has to be chosen to be small enough to avoid unnecessary computations of the cost function. These two issues are effectively and efficiently solved by performing the integer search with functions that bound the original cost function $C(\mathbf{Z})$ and are computationally easier to evaluate. Formally, two integer search procedures can be devised, depending on whether one works with lower or upper bounds of function $C(\mathbf{Z})$.

Search-and-Expand Algorithm. Let $C_l(\mathbf{Z}) \leq C(\mathbf{Z})$ be a suitable lower bound of $C(\mathbf{Z})$. The search space

associated to $C_1(\mathbf{Z})$, defined as

$$S_1(\chi^2) = \{\mathbf{Z} \in \mathbb{Z}^{n \times m} \mid C_1(\mathbf{Z}) \leq \chi^2\} \quad (27.57)$$

is a superset of the original search space (27.56)

$$S(\chi^2) \subseteq S_1(\chi^2). \quad (27.58)$$

First, the search space $S_1(\chi_0^2)$, with initial value χ_0 chosen to be arbitrarily small, is explored aiming for integer candidates within its boundaries. If a nonempty set of integer candidate is found, function $C(\mathbf{Z})$ is evaluated for each element of the set, thus enumerating integer candidates within the original search space $S(\chi_0^2)$. Should set $S(\chi_0^2)$ be nonempty, the potential candidates \mathbf{Z}_i are sorted according to the value of $C(\mathbf{Z}_i)$, and the candidate that returns the smallest value of the cost function is the sought integer minimizer $\check{\mathbf{Z}}$. If set $S(\chi_0^2)$ is empty, the search iteratively continues with the same modalities, by increasing the scalar $\chi_{s+1}^2 > \chi_s^2$ at each loop.

Search-and-Shrink Algorithm. Let $C_u(\mathbf{Z}) \geq C(\mathbf{Z})$ be an upper bound of $C(\mathbf{Z})$, with associated search space

$$S_u(\chi^2) = \{\mathbf{Z} \in \mathbb{Z}^{n \times m} \mid C_u(\mathbf{Z}) \leq \chi^2\}, \quad (27.59)$$

where $S_u(\chi^2) \subseteq S(\chi^2)$. First, an initial value for the scalar $\chi^2 = \chi_0^2$ is chosen such to guarantee the nonemptiness of $S_u(\chi_0^2)$. As soon as an integer matrix $\mathbf{Z}_1 \in S_u(\chi_0^2)$ is found, such that $C_u(\mathbf{Z}_1) < \chi_0^2$, the search space is shrunk by replacing $\chi_1^2 = C_u(\mathbf{Z}_1)$, and the search continues in the shrunken set $S_u(\chi_1^2)$, looking for an integer candidate \mathbf{Z}_2 that returns $C_u(\mathbf{Z}_2) < \chi_1^2$. The search proceeds iteratively until the minimizer of $C_u(\mathbf{Z})$, denoted with $\check{\mathbf{Z}}_u$, is found. Since $C_u(\check{\mathbf{Z}}_u) \leq C(\check{\mathbf{Z}}_u)$, the integer matrix $\check{\mathbf{Z}}_u$ may differ from the sought minimizer $\check{\mathbf{Z}}$. Thus, a final search is performed within the search space $S(\bar{\chi}^2)$, with $\bar{\chi}^2 = C_u(\check{\mathbf{Z}}_u)$. This final search is performed in a largely shrunken set, and only a few candidates need to be tested in order to extract the integer minimizer of $C(\mathbf{Z})$ [27.70].

The advantages of the constrained integer search algorithms described above are twofold. First, the size of the search space is iteratively adjusted during the search, avoiding the computation of the cost function for a large number of integer candidates. Second, the bounds used can be chosen among a class of *easy-to-compute* functions, avoiding the computational complexity of $C(\mathbf{Z})$. An example of lower and upper bounds is obtained by taking the smallest (ξ_{\min}) and the largest

(ξ_{\max}) eigenvalues of the inverse of matrix $\mathbf{Q}_{\hat{\mathbf{R}}(\mathbf{Z})\hat{\mathbf{R}}(\mathbf{Z})}$, which yield the following inequalities

$$\begin{aligned} \xi_{\min} \left\| \text{vec}(\hat{\mathbf{R}}(\mathbf{Z}) - \check{\mathbf{R}}(\mathbf{Z})) \right\|_{\mathbf{I}}^2 &\leq \left\| \text{vec}(\hat{\mathbf{R}}(\mathbf{Z}) - \check{\mathbf{R}}(\mathbf{Z})) \right\|_{\mathbf{Q}_{\hat{\mathbf{R}}(\mathbf{Z})\hat{\mathbf{R}}(\mathbf{Z})}}^2 \\ &\leq \xi_{\max} \left\| \text{vec}(\hat{\mathbf{R}}(\mathbf{Z}) - \check{\mathbf{R}}(\mathbf{Z})) \right\|_{\mathbf{I}}^2. \end{aligned} \quad (27.60)$$

The first and last squared norms in (27.60) are decomposed as

$$\sum_{i=1}^p \|\hat{\mathbf{r}}_i(\mathbf{Z})\|^2 + 1 - 2 \|\hat{\mathbf{r}}_i(\mathbf{Z})\| \cos(\alpha_i),$$

with α_i the angle formed by vectors $\hat{\mathbf{r}}_i(\mathbf{Z})$ and $\check{\mathbf{r}}_i(\mathbf{Z})$, which denote the i -th columns of $\hat{\mathbf{R}}$ and $\check{\mathbf{R}}$, respectively. Thus, two bounds $C_l(\mathbf{Z})$ and $C_u(\mathbf{Z})$ are defined as

$$\begin{aligned} C_l(\mathbf{Z}) &= \left\| \text{vec}(\hat{\mathbf{Z}} - \mathbf{Z}) \right\|_{\mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{Z}}}}^2 \\ &\quad + \xi_{\min} \sum_{i=1}^p (\|\hat{\mathbf{r}}_i(\mathbf{Z})\| - 1)^2, \\ C_u(\mathbf{Z}) &= \left\| \text{vec}(\hat{\mathbf{Z}} - \mathbf{Z}) \right\|_{\mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{Z}}}}^2 \\ &\quad + \xi_{\max} \sum_{i=1}^p (\|\hat{\mathbf{r}}_i(\mathbf{Z})\| - 1)^2. \end{aligned} \quad (27.61)$$

The evaluation of these functions does not require the solution of the constrained least-squares problem in (27.55), but only the computation of squared norms, thereby reducing the computational burden.

The choice of the bounds follows two criteria. First, the evaluation of the bound should be numerically efficient, avoiding nonlinear estimation problems. Second, the bounds should be tight enough with respect to the original function $C(\mathbf{Z})$ to guarantee fast convergence to the sought integer minimizer. Examples of efficient alternative bounds can be found in [27.71].

27.4.4 The Quality of Ambiguity and Attitude Estimations

In the first step of the attitude-ambiguity estimation procedure, float estimators are derived as linear functions of the observables. Thus, the float estimations distribute as

$$\begin{pmatrix} \text{vec}(\hat{\mathbf{Z}}) \\ \text{vec}(\hat{\mathbf{R}}) \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \text{vec}(\mathbf{Z}) \\ \text{vec}(\mathbf{R}) \end{pmatrix}, \begin{bmatrix} \mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{Z}}} & \mathbf{Q}_{\hat{\mathbf{Z}}\hat{\mathbf{R}}} \\ \mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{Z}}} & \mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}} \end{bmatrix} \right). \quad (27.62)$$

Table 27.1 10^5 data samples simulation of the single-baseline ($p = 1$), single-epoch, single-frequency ($f = 1$) success rates for the LAMBDA and MC-LAMBDA methods. The success rates are given per number of channels, and per code (pr) and phase noise (cp) levels combinations

σ_{cp} (mm)	3			1		
σ_{pr} (cm)	30	15	5	30	15	5
# Chan	LAMBDA					
	MC-LAMBDA					
5	0.03	0.19	0.87	0.60	0.27	0.95
	0.72	0.89	0.99	0.97	0.99	1.00
6	0.25	0.67	0.97	0.49	0.87	0.99
	0.96	0.99	0.99	0.99	1.00	1.00
7	0.50	0.79	0.99	0.74	0.93	1.00
	0.99	0.99	1.00	1.00	1.00	1.00
8	0.86	0.95	0.99	0.99	0.99	1.00
	0.99	0.99	1.00	1.00	1.00	1.00

In order to highlight the single terms that contribute to the precision of the float estimation, the v-c matrix of the observations is assumed to be decomposed as $\mathbf{Q}_{YY} = \mathbf{P}_m \otimes \mathbf{Q}_{yy}$. This enables the decomposition of the float ambiguity and attitude v-c matrices in (27.62) as

$$\begin{aligned}\mathbf{Q}_{\hat{Z}\hat{Z}} &= \left[\mathbf{P}_m^{-1} \mathbf{P}_{F^\perp}^\perp \otimes \tilde{\mathbf{A}}^\top \mathbf{Q}_{yy}^{-1} \tilde{\mathbf{A}} \right. \\ &\quad \left. + \mathbf{P}_m^{-1} \otimes \tilde{\mathbf{A}}^\top \mathbf{Q}_{yy}^{-1} \tilde{\mathbf{A}} \right]^{-1}, \\ \mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}} &= [\mathbf{F} \mathbf{P}_m^{-1} \mathbf{F}^\top]^{-1} \otimes [\tilde{\mathbf{G}}^\top \mathbf{Q}_{yy}^{-1} \tilde{\mathbf{G}}]^{-1},\end{aligned}\quad (27.63)$$

with $\tilde{\mathbf{A}} = \mathbf{P}_G \mathbf{A}$ and $\tilde{\mathbf{A}} = \mathbf{P}_G^\perp \mathbf{A}$. The precision of both float estimators $\hat{\mathbf{Z}}$ and $\hat{\mathbf{R}}$ depends on the number of satellites and frequencies tracked, on the satellite geometry (\mathbf{G}), and on the quality of the observations (\mathbf{Q}_{yy}). In details, it is the precision of the pseudorange observations that influences the float estimation error [27.32, 69]. The quality of the float estimation also depends on the mutual geometry of the baselines formed by the elements of the antenna array. By increasing the size of the antenna array \mathbf{F} with a scaling factor g , the elements of the float attitude v-c matrix reduce by a factor $1/g^2$, thus capturing the impact of longer baselines. Also the angles formed by the baselines play a role in defining the entries of $\mathbf{Q}_{\hat{\mathbf{R}}\hat{\mathbf{R}}}$, although this dependency is somewhat more complicated to figure. As a general rule, a balanced baseline configuration, intended as a distribution of baselines directed so to achieve maximum angular separations, should be preferred for a more precise estimation of the whole set of attitude parameters [27.72]. On the contrary, multiple aligned baselines may be used to enhance the estimation of rotations about a particular axis.

Differently from float attitude estimates, the precision of the float ambiguities is independent of any

Table 27.2 10^5 data samples simulation of the two-baseline ($p = 2$, noncollinear), single-epoch, single-frequency ($f = 1$) success rates for the LAMBDA and MC-LAMBDA methods. The success rates are given per number of channels, and per code (pr) and phase noise (cp) levels combinations

σ_{cp} (mm)	3			1		
σ_{pr} (cm)	30	15	5	30	15	5
# Chan	LAMBDA					
	MC-LAMBDA					
5	0.01	0.06	0.84	0.01	0.10	0.96
	0.99	0.99	1.00	1.00	1.00	1.00
6	0.10	0.57	0.97	0.30	0.81	1.00
	0.99	1.00	1.00	1.00	1.00	1.00
7	0.32	0.73	0.99	0.61	0.91	1.00
	0.99	1.00	1.00	1.00	1.00	1.00
8	0.82	0.93	0.99	0.99	1.00	1.00
	1.00	1.00	1.00	1.00	1.00	1.00

arbitrary scaling of the antenna array. However, the ambiguity float estimation does benefit from certain geometries of the GNSS antenna array. The orthogonal projector $\mathbf{P}_{F^\perp}^\perp$ in (27.63) differs from zero for any baseline arrangement characterized by $p < m$ (Sect. 27.4). Thus, lower entries of the ambiguity float v-c matrix are obtained when employing:

1. Three or more collinear antennas
2. Four or more coplanar antennas
3. Any configuration of five or more antennas [27.35, 73].

The reason why these configurations yield improved float ambiguity estimation is the presence of implicit linear constraints in the GNSS attitude model (27.45), as discussed in Sect. 27.4.

An analysis of the quality of the constrained ambiguity-attitude estimation is somewhat more complex. As with the unconstrained ILS, the precision of the ambiguity estimation depends on the quality of the float estimation. However, the latter does not completely define the success rate of the constrained approach, due to the additional geometrical constraints in (27.54). By estimating the integer ambiguities and the platform's attitude in an integral manner, fully exploiting the known body geometry of the multiantenna configuration, the ambiguity resolution performance is greatly improved with respect to unconstrained integer ambiguity resolution methods [27.67, 69]. The enhancement of the ambiguity resolution performance can be appreciated through simulations, which provide a comparison between the performance of unconstrained and constrained ILS estimation in a controlled environment. Tables 27.1–27.2 give the success rate,

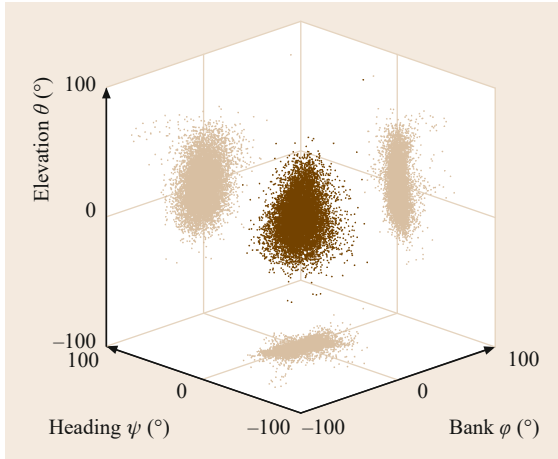


Fig. 27.10 Float attitude angles, estimated from 9000 GPS data epochs from nine satellites tracked by two static baselines, each 2 m long

defined as the ratio of samples in which the correct integer ambiguity matrix is found, as function of number of measurement channels (n) and observation noise.

The results from two single-epoch, single-frequency ($f = 1$) scenarios are given, for a single-baseline ($p = m = 1$) and a two-baseline ($p = m = 2$) antenna array, with local baseline coordinate matrices.

The improvement in terms of successful ambiguity resolution when applying the constrained ambiguity-attitude estimation procedures is rather large, particularly for the weaker scenarios, that is, with fewer measurements and higher observation noise. Increasing the number of baselines has a very important impact on the constrained success rate, with success rates equal or close to one on all the simulated scenarios. As expected, the inclusion of additional constraints largely strengthens the GNSS observation model, and substantially affects the ambiguity success rate, even for an array formed by a limited number of antennas (two or three in the examples reported).

$$\mathbf{F}_{p=1} = 1 \quad ; \quad \mathbf{F}_{p=2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \quad (27.64)$$

Assuming the carrier-phase ambiguities as fixed to their correct integer values \mathbf{Z} , the conditional attitude solution is obtained as in (27.53). The precision of the conditional attitude estimation would then be described by a normal distribution with v-c matrix $\mathbf{Q}_{\hat{\mathbf{R}}(\mathbf{Z})\hat{\mathbf{R}}(\mathbf{Z})}$. However, the statistics of the integer estimation must also be taken into account. The integer ambiguities distribute according to a probability mass function (PMF), obtained by integrating the normal distribution of the float estimation $\hat{\mathbf{Z}}$ over the pull-in regions associated to

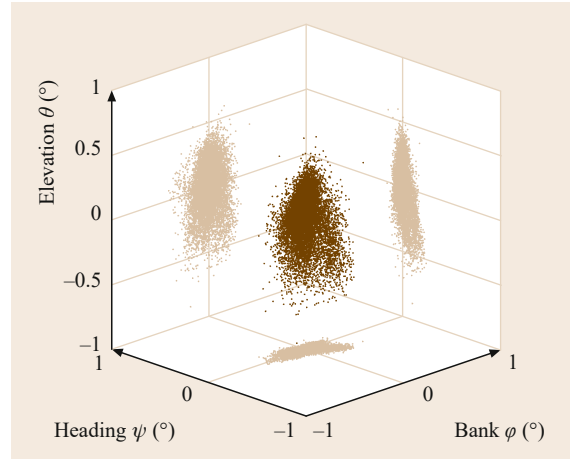


Fig. 27.11 Attitude angles after successful ambiguity resolution, estimated from 9000 GPS data epochs from nine satellites simultaneously tracked by two static baselines, each 2 m long

the ambiguity-attitude estimator. The result is a multimodal distribution of type [27.74]

$$f_{\hat{\mathbf{R}}(\hat{\mathbf{Z}})}(x) = \sum_{\mathbf{N} \in \mathbb{Z}^{n \times m}} f_{\hat{\mathbf{R}}(\mathbf{N})}(x) P(\hat{\mathbf{Z}} = \mathbf{N}), \quad (27.65)$$

where $f_{\hat{\mathbf{R}}(\mathbf{N})}(x)$ denotes the probability density function of the conditional attitude estimation and $P(\hat{\mathbf{Z}} = \mathbf{N})$ denotes the PMF of the fixed ambiguities evaluated at \mathbf{N} . When the success rate $P_s = P(\hat{\mathbf{Z}} = \mathbf{Z})$ is sufficiently close to the unit, the distribution of the conditional attitude solution is well approximated by the normal distribution with v-c matrix $\mathbf{Q}_{\hat{\mathbf{R}}(\mathbf{Z})\hat{\mathbf{R}}(\mathbf{Z})}$, which provides a first-order approximation of the dispersion of the attitude estimator $\hat{\mathbf{R}}$. The precision of the attitude solution is driven by the precision of the carrier-phase observables [27.15] and by the geometry of the antenna array, for which the same considerations drawn for the attitude float solution apply: longer baselines should be employed to achieve smaller angle estimation errors, compatibly with structural and physical limitations imposed by the object or platform whose attitude has to be measured.

The large improvement in terms of angular estimation accuracy after successful ambiguity resolution is clear when comparing Fig. 27.10 with Fig. 27.11, in which the attitude angles before and after ambiguity resolution are shown for a two-baseline static array, where each baseline is 2 m long. The precision of each of the attitude angles improves by two orders of magnitudes when the carrier-phase ambiguities are successfully fixed to the correct integer values (Table 27.3).

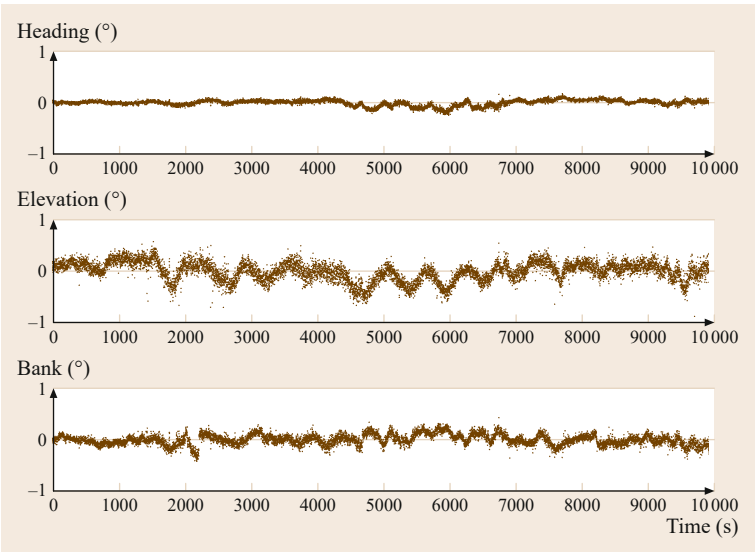


Fig. 27.12 Time series of the three attitude angles after successful ambiguity resolution, 9000 GPS data epochs from nine satellites tracked by two static baselines, each 2 m long

Table 27.3 Comparison between the precision of attitude estimation before and after successful ambiguity resolution. **RMS:** root mean square

Component	Float	Fixed
Heading (°), RMS	6.35	0.05
Elevation (°), RMS	16.01	0.19
Bank (°), RMS	12.51	0.10

The differences among the estimation errors for the three orientation angles highlight a common property of GNSS-based attitude sensors (Table 27.3). The two baselines of the static array used in the previous test lie on a locally horizontal plane. Both elevation and bank angles are then estimated with a lower precision than the heading angle, due to the poorer geometric dilution of precision that characterizes the vertical component

of the observations. In general, one may expect better angular estimations for rotations about the direction of the line-of-sight vectors to the navigation satellites.

Figure 27.12 shows the heading, elevation and bank angles as output epoch-by-epoch by the three-antenna GNSS attitude sensor after the integer ambiguities are correctly resolved. The effect of multipath is clearly visible. This shows as a varying, time-correlated error in all three angular estimates, superimposed to the Gaussian-like estimation noise. The magnitude of the bias varies with time, with peaks reaching in the worst case half a degree, in agreement with a centimeter-level multipath error on the two-meter-long baselines. The error introduced by multipath is also differently amplified in each of the three angular estimates, with the heading output being the least affected.

27.5 Applications

27.5.1 Space Operations

The use of multiple GNSS antennas to keep track of the orientation in spacecraft and space structures was investigated at a very early stage of the satellite navigation era. Early in the 1990s, a Trimble Navigation TANS (acronym of Trimble Advanced Navigation Systems) Quadrex [27.39, 75], fittingly modified by Stanford University, was flown in the National Aeronautics and Space Administration (NASA) space mission RADCAL (Radar Calibration, 1993) [27.76]. The RADCAL mission first tested the performance of spacecraft attitude determination via the GPS, although only in post-

processing, achieving 1° (root mean squared [RMS]) attitude precision with four half-meter baseline lengths. In 1994 the CRISTA-SPAS (Cryogenic Infrared Spectrometers and Telescopes for the Atmosphere – Shuttle Palette Satellite) mission [27.77], designed and developed by the University of Wuppertal to measure infrared emissions of the Earth’s atmosphere, demonstrated the capabilities of real-time GPS-based attitude determination (Fig. 27.13). Shortly after, an updated version of Quadrex GPS receiver, the Trimble Quad Vector, was developed and used on board the NASA mission APEX (Advanced Photovoltaic and Electronic Experiments) [27.78], and on several shuttle flights [27.77].

In 1996 the REX-II (Radiation Experiment Satellite, second mission) satellite, equipped with an updated TANS receiver, achieved the first successful implementation of a real-time GPS-based attitude estimation for closed-loop control. In the REX-II spacecraft the combination of carrier-phase readings from four GPS antennas tracking at most six satellites were coupled to magnetic field measurements, fulfilling the mission requirements of 2° (RMS) angular accuracy. The performance of a standalone GPS attitude sensor was also evaluated at 5° (RMS). An issue related to the GPS attitude sensor was availability, with frequent dropouts caused by insufficient number of satellites, poor satellite geometry, and noisy signals [27.79].

GNSS-based attitude estimation has been implemented in a number of space missions, such as, for example, the GPS Attitude/Navigation Experiment (GANE), the GPS-based Meteorology (GPS/MET), the SSTL satellites UoSat-12 and the Tactical Operational Satellite (TopSat) [27.80, 81]. Notably, an array of GNSS antennas was also carried on the Gravity Probe-B Relativity Mission, whose data were used to support the post-processing necessary to validate two predictions of Einstein's theory of general relativity: the geodetic and frame-dragging effects [27.82].

A four-antenna attitude system is currently operational on board the International Space Station (ISS) (Fig. 27.14). The system provides attitude information that fulfills the requirements (set to 0.5° error, three sigma) only in combination with the onboard inertial navigation system (INS) [27.83]. The main issue of a GNSS array installed on large space structures is the potentially large multipath environment, created by both the craft structures and the solar panels. Satellite

blockages may also occur, due to obstruction from the structure itself, from mobile appendices (such as the robotic arm and deployed solar panels on the ISS), and eventually from docked vehicles.

Despite the lower precision, GNSS attitude systems on board spacecraft retain several advantages when compared to horizon sensors and star trackers, principally in terms of maintenance, cost and weight. These advantages make GNSS attitude systems a viable alternative to other sensors for small satellites flying in low Earth orbits (LEOs) when the accuracy requirements are not too stringent, as the precision of angular estimates is usually limited by the small size of the space platform.

The availability of attitude sensing depends on the actual spacecraft rotational motion: a GNSS attitude sensor on spinning satellites may experience frequent satellite losses and reacquisitions. This aspect could be mitigated by the large number of navigation satellites expected to be operational in the near future.

Space application of GNSS attitude sensors also include support of rendezvous [27.84] and reentry maneuvering. However, the latter application has been so far rejected in favor of a combination of inertial measurement units (IMUs) and star trackers. The limiting factor of a GNSS attitude system during the Earth reentry phases is satellite visibility: the typical flight envelope is characterized by several rotations of the platform, and adequate visibility cannot be guaranteed [27.85].

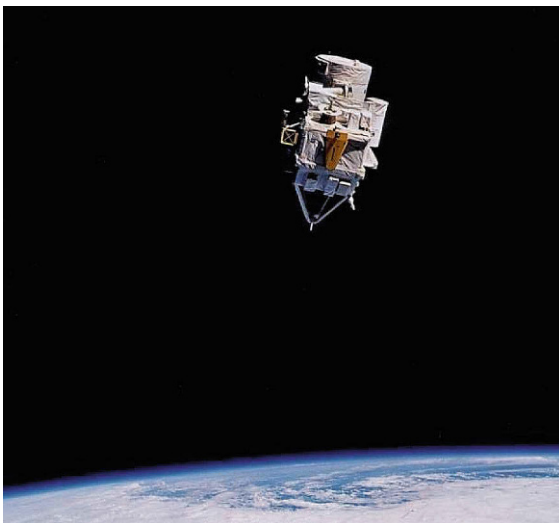


Fig. 27.13 The CRISTA-SPAS 2 flight. Credit: NASA

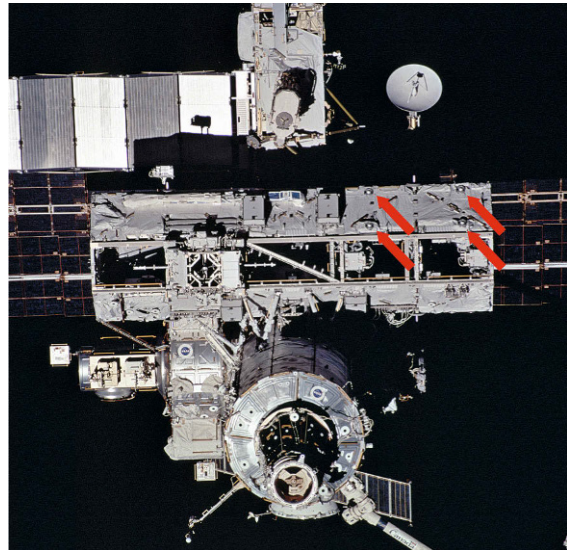


Fig. 27.14 The International Space Station as seen from the shuttle Atlantis, STS-110. Four GNSS choke-ring antennas (marked by red arrows) are mounted on the S0 Truss and are visible on the right side of the picture. Credit: NASA

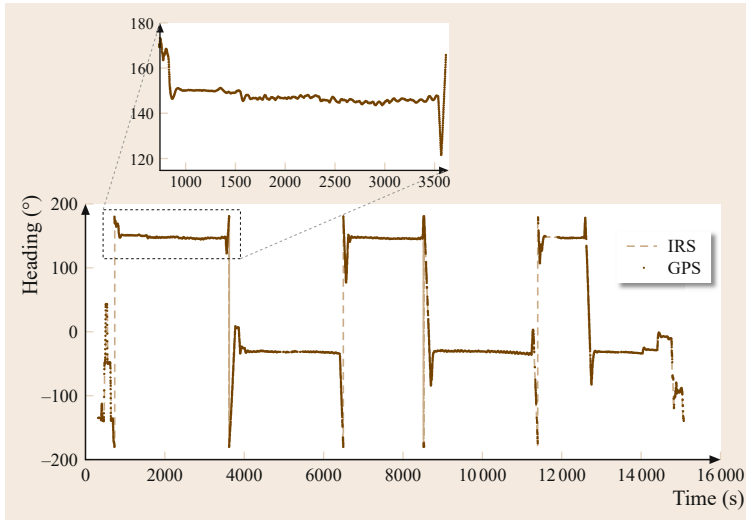


Fig. 27.15 Epoch-by-epoch heading angle estimates from a three-antenna GNSS attitude system and an inertial reference system, during a four-hour flight

27.5.2 Aeronautics Applications

Starting from the 1990s, with the development of more powerful, smaller and less expensive receivers, GNSS antenna arrays as orientation sensors became widely used in aeronautics applications, either as standalone systems or, more often, integrated with IMUs.

GNSS attitude sensors are an effective navigation aid for the guidance of civil and military airborne vehicles in all the phases of flight: takeoff, en-route, formation flying, landing and taxiing [27.50, 86–89]. A multiple-GNSS antenna system is often integrated in the autonomous guidance control system of unmanned airborne vehicles (UAVs) [27.90], which take advantage of the low-cost, power-efficient and driftless properties of GNSS attitude sensors. In addition, GNSS attitude information is a valuable input for georeferencing, that is, the post processing of data acquired remotely from aircraft or other flying platforms [27.91, 92].

The use of a GNSS multiantenna system provides attitude information that remains unaffected by drift and magnetic variations. Also, the typical size of an aircraft, with fuselage and wingspan longer than a few meters, contributes to enhance the GNSS-based angular estimation precision. An example is shown in Fig. 27.15, which reports the heading angle provided epoch-by-epoch by a standalone three-antenna GNSS attitude system. The three antennas are located on the nose, on the left wing and on the central section of the fuselage of a Cessna Citation II, forming the following matrix of local baseline coordinates

$$\mathbf{F} = \begin{bmatrix} 4.90 & -0.39 \\ 0 & 7.60 \end{bmatrix} [\text{m}]. \quad (27.66)$$

The precise heading angle provided from an inertial reference system (IRS), a *Honeywell LaseRef*, is also shown in Fig. 27.15: the difference between the GNSS and IRS output amounts to 0.07° (RMS). The cumulative error between the GNSS-based attitude output and the IRS for the three attitude angles (heading, pitch and roll) is shown in Fig. 27.16. Of the three angles, the heading is the most accurate, whereas the pitch angle shows the noisiest behavior. The RMS of the IRS-to-GNSS output difference is 0.12° for the roll angle and 0.3° for the pitch angle. These values, related to the baseline lengths used, are typical for a GNSS attitude system operating on airborne structures. The accuracy of GNSS-based attitude estimates is limited by the combination of two main factors: multipath and structural flexibility.

Multipath usually has an important impact on the accuracy of the angular estimates in aeronautics applications due the siting of the GNSS antennas. These must be placed on the metallic surface of an aircraft without significantly altering its aerodynamic properties, thereby imposing limitations on the shape and size of the antennas employed. The primary effect of these limitations are measurements of somewhat lesser quality, with two consequences: potentially wrong ambiguity resolution and larger angular estimation errors and/or biases. These effects are clearly shown in Fig. 27.17, which reports the carrier-phase estimation residuals for a portion of the four-hour flight examined above: the residuals are noisy and for the most part biased. The culprit is a disturbed signal received at the nose antenna from a satellite at low elevation and azimuth opposite to the flight direction. Correctly identifying such degraded measurements is of primary impor-

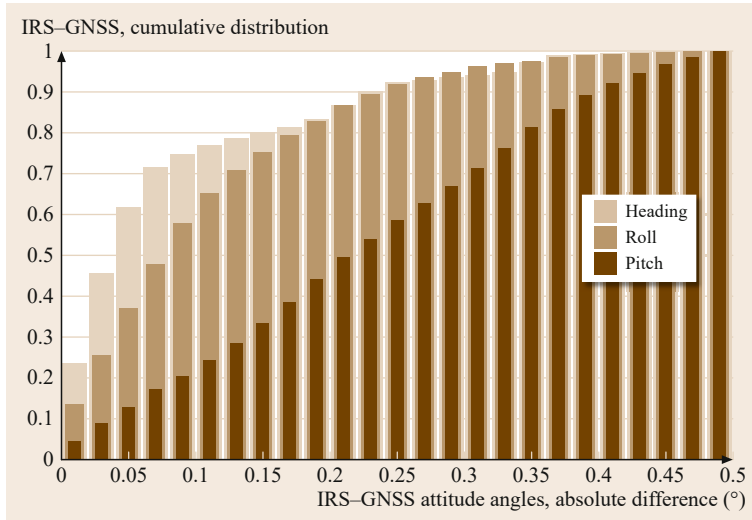


Fig. 27.16 Cumulative distribution of the IRS-GNSS attitude angle output differences during a four-hour flight

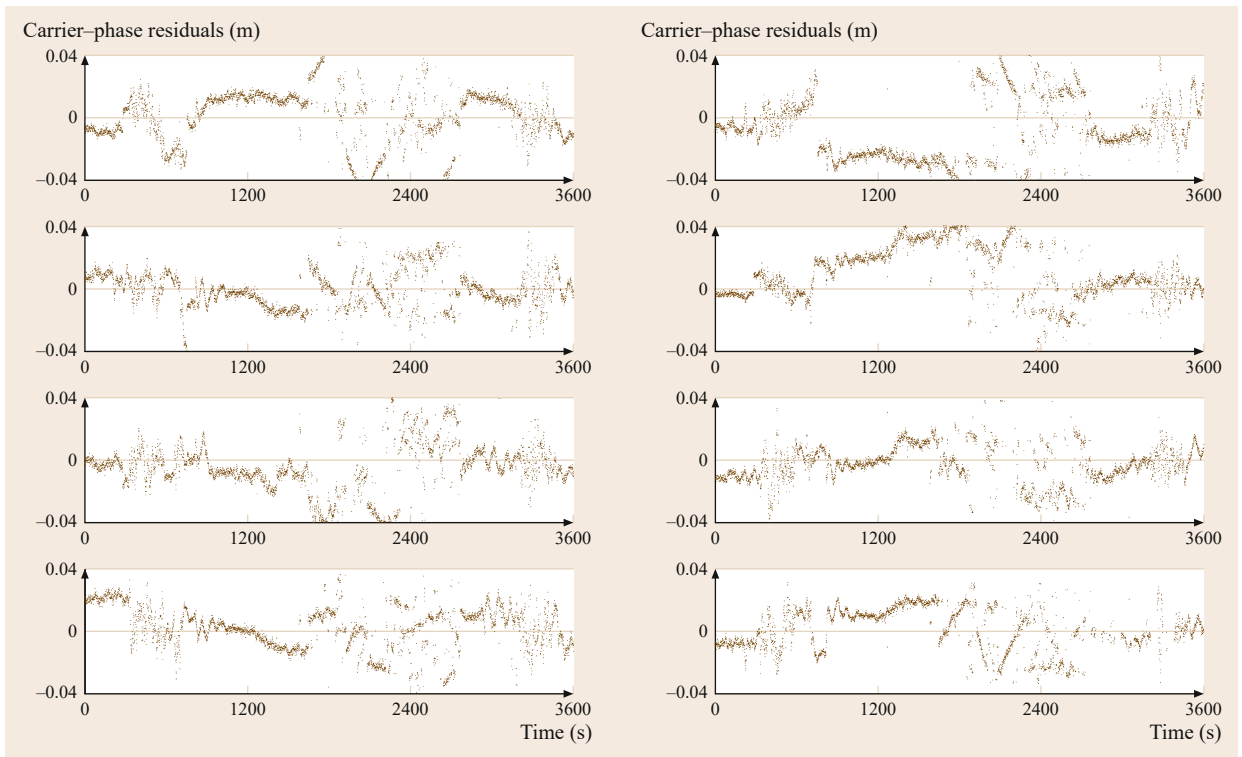


Fig. 27.17 Carrier-phase estimation residuals prior to elimination of a disturbed signal, during a portion of a four-hour flight

tance to avoid estimation biases: Fig. 27.18 shows the carrier-phase residuals after eliminating the disturbed signal. After correction, the time characteristic of the residuals only shows a small remaining time-correlated error, superimposed to the expected millimeter-level carrier-phase measurement noise.

The second, potentially large error source in aeronautics applications is the platform flexibility. In order to fully exploit the size of an aircraft, one (or more) antennas should be mounted as close to the wingtip as possible. However, such placement would likely introduce errors caused by misalignment of the baseline

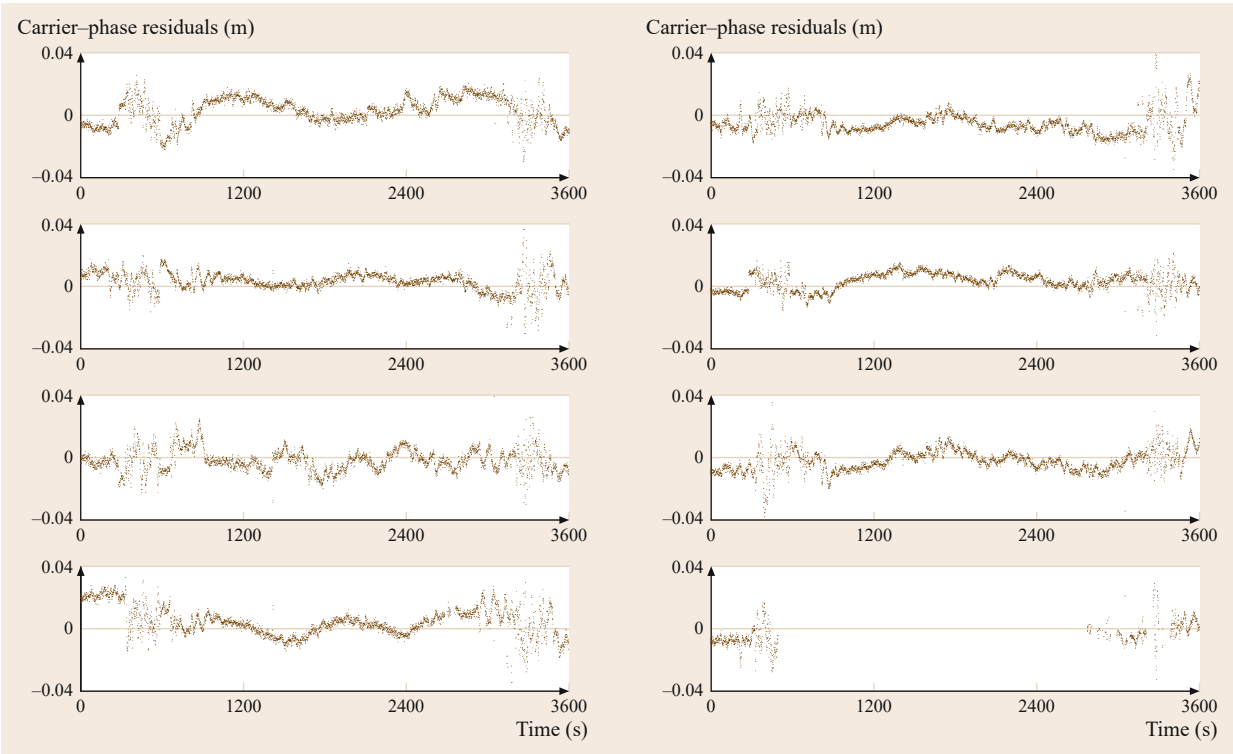


Fig. 27.18 Carrier-phase estimation residuals after the elimination of a disturbed signal, during a portion of a four-hour flight

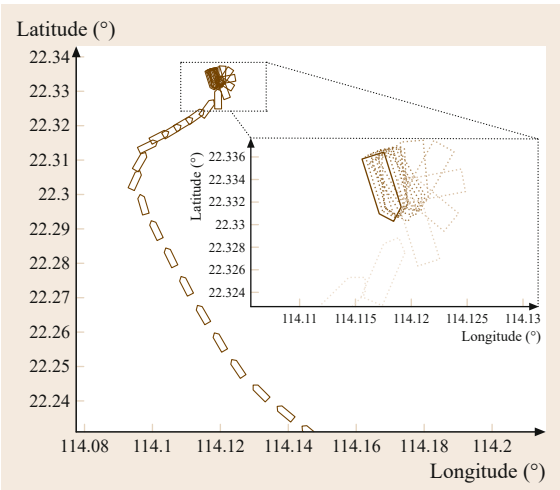


Fig. 27.19 The heading of a large container ship being finely tracked by a GNSS attitude sensor during its final approach to the quay

coordinates in the local frame, due to deformations caused by maneuvering, turbulence, and fuel consumption, all of which alter the wing load. If these changes are not correctly modeled, estimation errors and biases may occur, degrading the GNSS attitude output.

27.5.3 Marine Navigation

Among all possible applications of GNSS as attitude sensor, marine navigation is arguably the one that may benefit the most from GNSS orientation guidance. Firstly, navigation on waterways usually offers a clear-sky, low-dynamic environment, in which all the signals can be received without excessive degradation. Secondly, antennas of any type can be installed on boats such to form baselines ranging from a few meters to several hundreds of meters, reducing multipath from the boat’s own structures and/or load through the use of antenna masts.

GNSS antenna arrays are currently used on ships of different classes to provide either heading estimation and full three-axis orientation, aiding the guidance during navigation in open waters [27.93–98], enhancing the ship attitude monitoring when navigating in shallow waters in order to control the underkeel clearance and perform grounding avoidance [27.47, 99], and during berthing maneuvers [27.100].

An example of the latter application is given in Fig. 27.19, where the heading of a large container ship is shown during its final approach to the quay at the Hong Kong harbor. The heading is extracted from a three-antenna array system with two antennas on the

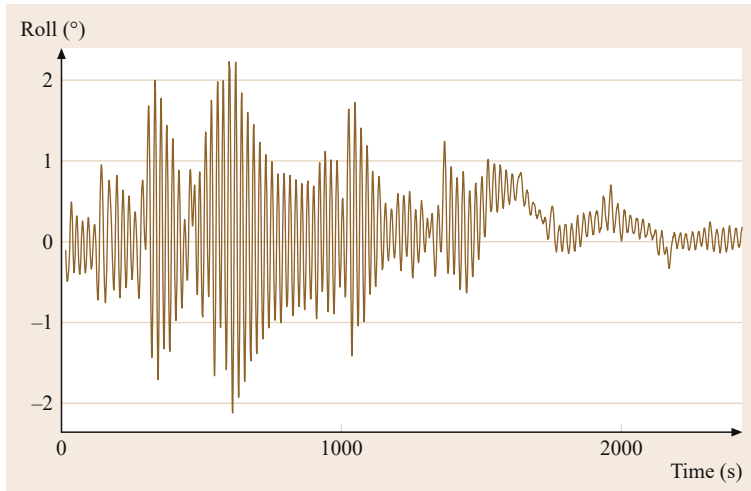


Fig. 27.20 Tracking of a vessel's rolling during the transition from open water to the calmer harbor area

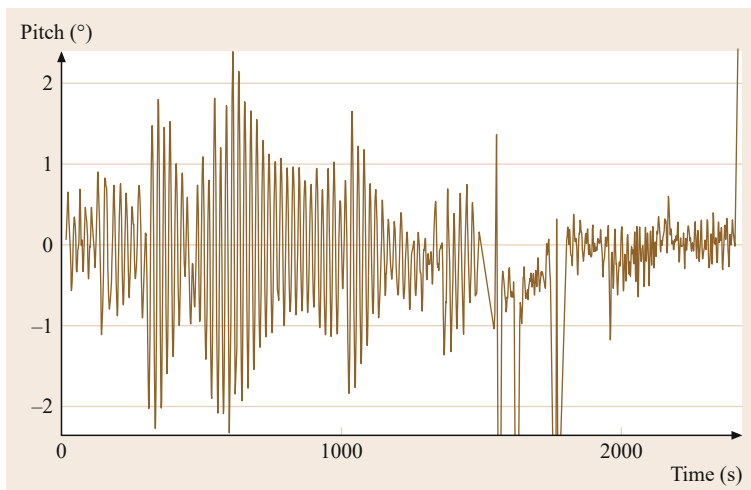


Fig. 27.21 Tracking of a vessel's pitching during the transition from open water to the calmer harbor area

bridge – one at port and one at starboard side, distant 40 m – and one antenna at the bow, at 213 m from each of the antennas on the bridge. The very long baselines employed guarantee heading estimation precision in the region of 10^{-2} – 10^{-3} degrees. Only slightly lower precision is achieved in the estimation of the roll and pitch angles. These are given in Figs. 27.20 and 27.21, which track the vessel attitude during its entrance in the harbor area. Both pitching and rolling reduce considerably during the transition, and both movements can be perfectly tracked by the GNSS attitude sensor.

Two error sources are most likely to have an impact on GNSS attitude systems on board vessels. The first is again multipath, which can be reduced by proper antenna siting and employing ad hoc antennas such as choke rings. The second, predominant on large vessels, is the boat structural deformation. This can be rather large on ships whose length exceeds more than a few

tens of meters, and cannot be easily decoupled from actual rotations without an external aid or proper filtering. The effect of longitudinal deformations can be reduced by designing baseline lengths that maximize the accuracy without amplifying the estimation biases caused by mismodeling of the local body frame.

27.5.4 Land Applications

GNSS-based attitude determination is extensively used in the realm of land navigation. Multiantenna GNSS systems are being adopted in the automotive sector for the determination of heading, roll and pitch angles, to derive slip angle measurements, or to assist lorry drivers in maneuvering the truck and trailer [27.51, 101, 102]. Figure 27.22 shows a portable digital heading gage implemented on a tablet, which can be fed by low-cost GNSS antennas mounted on a vehicle rooftop.



Fig. 27.22 Low-cost GNSS-based digital heading gage and tracking system developed for automotive applications, ANavS, Munich. Credit: ANavS

Although GNSS attitude systems are capable of meeting the accuracy required in most automotive applications, the availability and continuity of the attitude output still represents an issue, due to signal obstructions from tunnels or surrounding structures (e.g., urban canyons), thereby requiring integration with other navigation means.

GNSS antenna arrays in combination with inertial guidance are often a primary navigation instrument for unmanned ground vehicles (UGVs), ranging from large agricultural machinery (precision farm-

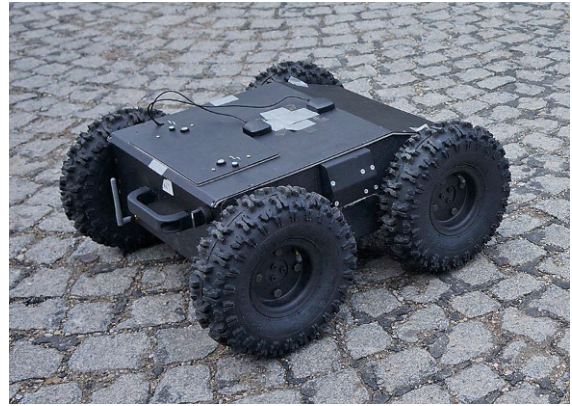


Fig. 27.23 Example of an unmanned ground vehicle equipped with a two-antenna GNSS heading system

ing) [27.103] to small robotic platforms [27.104, 105], as in Fig. 27.23.

Further potential uses of GNSS attitude determination include heading systems for trains [27.106], support in the orientation control of large structures such as antennas or antenna arrays, and tracking of the mobile arm of cranes. Furthermore, GNSS antenna arrays are of common use in remote sensing campaigns, providing an attitude reference for georeferencing of sensed data [27.107–110].

27.6 An Overview of GNSS/INS Sensor Fusion for Attitude Determination

Because of the complementary advantages of each system, GNSSs and INSs are ideally suited for an integrated attitude solution, with the INS platform providing lower measurement noise, higher output rate and overall improved tracking robustness, and the GNSS supplying a stable reference solution for initialization and recalibrations.

INSs exploit self-contained inertial measurement units (IMUs) to sense the variations in the body position and orientation. In an inertial system, the angular motion is sensed with gyroscopes. These can be based on mechanical principles, such in gimballed gyros or gyrostats, which maintain angular momentum in the presence of rotational forces; on optical principles, such in fiber optic or laser gyros, based on the *Sagnac effect* [27.111] (rotation-induced perturbations on light beams interferometry); on quantum-mechanical phenomena, such as in London-moment gyroscopes [27.112]; and on electromechanic principles, such in MEMSs (Micro Electro-Mechanical Systems), which measure the Coriolis force generated from vi-

brating elements subject to rotations of the plane of vibration. The wide range of inertial orientation sensors differ in terms of cost, weight, structural complexity, and measurement sensitivity, but most are superior to GNSS-based attitude sensors in terms of (instantaneous) angular measurement precision and output rate. Also, INSs do not suffer from outages due to lack of external references caused by jamming or signal blocking. However, the common trait of most INSs is the derivation of the body orientation through integration of rotational accelerations, starting from a known initial state (dead reckoning). Thus, the INSs require an initialization step, followed by periodic recalibrations due to the lack of long-term stability.

Several GNSS/INS architectures can be realized, depending on the required navigation quality and robustness. The coupling of the two systems can be implemented at different levels of tightness [27.113, 114]:

- Loose integration, where the two systems are kept separated, providing two independent navigation

solutions that are subsequently filtered to output an enhanced common navigation solution. A feedback channel is usually designed to attenuate the INS drift.

- Tight integration, in which the coupling is implemented in the measurement domain, with GNSS pseudoranges, carrier-phase and/or Doppler observations linked to the INS output to enhance observability, reduce measurement noise and improve fault detectability [27.115].
- Ultratight integration, in which the coupling is realized already in the GNSS tracking loop, with improvements in GNSS tracking robustness, noise reduction, fault recovery and overall robustness with respect to the platform dynamics [27.116].

Any of the aforementioned architectures, exhaustively covered in Chap. 28, improves the quality of the integrated system with respect to the standalone solutions.

The fusion of GNSS and INS output is usually realized with linear or nonlinear filters, in which the measurement vector comprises the body angular displacement, velocity and/or acceleration. The rotational accelerations are sensed by the INS and integrated over time to extract attitude angles. These are complemented by the GNSS antenna array output, which provides the orientation angles of the reference system integral with the body with respect to the reference frame. The measurements are then processed by designing a proper fusion scheme for the nonlinear filter, where the body kinematics, the measurement error and the error propagation are modeled (Chap. 28). Four issues need to be addressed when realizing an integrated GNSS/INS solution: frame alignment, time synchronization and latency, GNSS integrity, and measurement fusion:

Frame alignment. An IMU senses the rotational accelerations and velocities around the inertial axes, that is, the axes (not necessarily orthogonal [27.117]) about which the gyro is capable of detecting rotations. The inertial axes do not necessarily coincide with the local body axes: linear (lever arm) and angular displacements between the two frames must be derived and compensated with an initial calibration (Chap. 28). Any misalignment would produce an erroneous interpretation of the INS angular output.

Time synchronization and latency. The IMU and the GNSS operate by sampling at different times, with a delay between measurements. This requires a synchronization scheme between the two systems. A coherent time reference can be achieved by exploiting the timing signal output by the GNSS receiver to time-reference the inertial measurements from the IMU [27.118]. Furthermore, the

latency of processing the GNSS raw measurements yields delayed delivery of the GNSS output to the filter. This may hinder real-time applications with high-dynamic platforms [27.119], and a latency compensation step must be considered in the designed fusion algorithm.

GNSS integrity. Outliers from the GNSS subsystem need to be detected and isolated to prevent large error propagations in the filter (Chap. 24). The task of monitoring the GNSS integrity is somewhat less complex in the integrated GNSS/INS approach than in the standalone case, as the inertial solution aids error and bias detection by providing a prediction of the GNSS measurements to be compared with the actual output.

Measurement fusion. The INS and the GNSS operate in different domains. The former samples angular accelerations and/or velocities around its own inertial axes, whereas the latter estimates the orientation of the body frame with respect to the reference frame. The two types of measurements are coupled by relating the INS output with the chosen set of attitude parameters. Assuming the inertial platform axes to be aligned with the local body frame \mathcal{F} , the angular parameters output by the INS need to be converted into rate of change of Euler angles or of quaternion of rotation. Denoting with $\boldsymbol{\omega}_I = (\omega_{I,1}, \omega_{I,2}, \omega_{I,3})^\top$ the vector of angular velocities obtained with an inertial sensor aligned with the body frame, the rate of variation of the Euler angles is obtained as [27.120, 121]

$$\begin{pmatrix} \dot{\psi}_I \\ \dot{\theta}_I \\ \dot{\phi}_I \end{pmatrix} = \mathbf{T} \boldsymbol{\omega}_I, \quad (27.67)$$

with \mathbf{T} a 3×3 transformation matrix associated to the Euler angle sequence chosen. The complete list of values assumed by the elements of \mathbf{T} for each sequence can be found in [27.120].

The rates of change of the quaternion elements are computed from the inertial angular velocity measurements as [27.121]

$$\begin{pmatrix} \dot{q}_0 \\ \dot{\mathbf{q}} \end{pmatrix} = \frac{1}{2} \begin{bmatrix} 0 & \omega_{I,3} & -\omega_{I,2} & \omega_{I,1} \\ -\omega_{I,3} & 0 & \omega_{I,1} & \omega_{I,2} \\ \omega_{I,2} & -\omega_{I,1} & 0 & \omega_{I,3} \\ -\omega_{I,1} & -\omega_{I,2} & -\omega_{I,3} & 0 \end{bmatrix} \begin{pmatrix} q_0 \\ \mathbf{q} \end{pmatrix}. \quad (27.68)$$

Relations (27.67)–(27.68) enable a coherent fusion between the angular measurements of INS and GNSS, aligned so as to refer to the same body reference system.

References

- 27.1 P. Axelrad, C.P. Behre: Satellite attitude determination based on GPS signal-to-noise ratio, *Proc. IEEE* **87**(1), 133–144 (1999)
- 27.2 V.W. Spinney: Application of the Global Positioning System as an attitude reference for near-Earth users, *ION Natl. Aerosp. Meet. New Front. Aerosp. Navig.*, Warminster (ION, Virginia 1976)
- 27.3 R.L. Greenspan, A.Y. Ng, J.M. Przyjemski, J.D. Veale: Positioning by interferometry with reconstructed carrier GPS: Experimental results, *Proc. 3rd Int. Geod. Symp. Satell. Doppler Position.*, Las Cruces (Physical Science Laboratory, Las Cruces 1982) pp. 1177–1198
- 27.4 A.K. Brown, T.P. Thorvaldsen, W.M. Bowles: Interferometric attitude determination using the Global Positioning System – A new gyrotheodolite, *Proc. 3rd Int. Geod. Symp. Satell. Doppler Position.*, Las Cruces (Physical Science Laboratory, Las Cruces 1982) pp. 1289–1302
- 27.5 K.M. Joseph, P.S. Deem: Precision orientation: A New GPS application, *Int. Telem. Conf.*, San Diego (1983)
- 27.6 W.S. Burgett, S.D. Roerman, P.W. Ward: The development and applications of GPS-determined attitude, *Natl. Telesyst. Conf. (NTC)*, San Francisco (IEEE, New York 1983)
- 27.7 L.R. Kruczynski, P.C. Li, A.G. Evans, B.R. Wermann: Using GPS to determine vehicle attitude: USS Yorktown test results, *Proc. ION GPS*, Colorado Springs (ION, Virginia 1989) pp. 163–171
- 27.8 G.H. Purcell Jr., J.M. Srinivasan, L.E. Young, S.J. Di Nardo, E.L. Hushbeck, T.K. Meehan Jr., T.N. Munson, T.P. Yuncck: Measurement of aircraft position, velocity, and attitude using rogue GPS receivers, *5th Int. Geod. Symp. Satell. Position.*, Las Cruces (Physical Science Laboratory, Las Cruces 1989)
- 27.9 F. van Graas, M. Braasch: GPS interferometric attitude and heading determination: Initial flight test results, *Navigation* **38**, 297–316 (1991)
- 27.10 A. Karger, J. Novak: *Space Kinematics and Lie Groups* (Routledge, New York 1985)
- 27.11 L. Euler: Formulae generales pro translatione quacunque corporum rigidorum (General formulas for the translation of arbitrary rigid bodies), *Novi Commentarii Academiae Scientiarum Petropolitanae* **20**, 189–207 (1776), in Latin
- 27.12 W.R. Hamilton: *Philos. Mag. J. Sci. (3rd Series)*, On quaternions; or on a new system of imaginaries in algebra, *Lond. Edinb. Vol. 3* (Taylor Francis, Dublin 1844) pp. 489–495
- 27.13 J.B. Kuipers: *Quaternions and Rotations Sequences* (Princeton Univ. Press, Princeton 1999)
- 27.14 M.D. Shuster: A survey of attitude representations, *J. Astronaut. Sci.* **41**(4), 439–517 (1993)
- 27.15 P.J.G. Teunissen: A general multivariate formulation of the multi-antenna GNSS attitude determination problem, *Artif. Satell.* **42**(2), 97–111 (2007)
- 27.16 J.R. Wertz: *Spacecraft Attitude Determination and Control*, 1st edn. (Kluwer Academic, Dordrecht 1978)
- 27.17 C.F. Van Loan: The ubiquitous Kronecker product, *J. Comput. Appl. Math.* **123**(1/2), 85–100 (2000)
- 27.18 G. Wahba: Problem 65-1: A least squares estimate of spacecraft attitude, *SIAM Rev.* **7**(3), 384–386 (1965)
- 27.19 P.H. Schönemann: A generalized solution of the orthogonal Procrustes problem, *Psychometrika* **31**(1), 1–10 (1966)
- 27.20 P.B. Davenport: *A Vector Approach to the Algebra of Rotations with Applications*, NASA Technical Note D-4696 (Goddard Space Flight Center, Greenbelt 1968)
- 27.21 D.W. Eggert, A. Lorusso, R.B. Fisher: Estimating 3-D rigid body transformations: A comparison of four major algorithms, *SIAM J. Matrix Anal. Appl.* **9**(5/6), 272–290 (1997)
- 27.22 M.D. Shuster, S.D. Oh: Three-axis attitude determination from vector observations, *AIAA J. Guid. Contr.* **4**(1), 70–77 (1981)
- 27.23 M.D. Shuster: The quest for better attitudes, *J. Astronaut. Sci.* **54**(3/4), 657–683 (2006)
- 27.24 F.L. Markley, F. Landis: Attitude determination using vector observations: A fast optimal matrix algorithm, *J. Astronaut. Sci.* **41**(2), 261–280 (1993)
- 27.25 D. Mortari: ESQ: A closed-form solution to the Wahba problem, *J. Astronaut. Sci.* **45**(2), 195–204 (1997)
- 27.26 D. Mortari: Second estimator of the optimal quaternion, *AIAA J. Guid. Contr. Dyn.* **23**(5), 885–888 (2000)
- 27.27 F.L. Markley, D. Mortari: How to estimate attitude from vector observations, *AAS 99-427, AAS/AIAA Astrodyn. Spec. Conf., Girdwood*, ed. by K.C. Howell, F.R. Hoots, B. Kaufman, K.T. Alfriend (Univelt, San Diego 1999) pp. 1979–1996
- 27.28 F.L. Markley, D. Mortari: Quaternion attitude estimation using vector observations, *J. Astronaut. Sci.* **48**(2/3), 359–380 (2000)
- 27.29 Y. Cheng, M.D. Shuster: Robustness and accuracy of the QUEST algorithm, *Adv. Astronaut. Sci.* **127**, 41–61 (2007)
- 27.30 M.T. Chu, N.T. Trendafilov: On a differential approach to the weighted orthogonal procrustes problem, *Stat. Comput.* **8**, 125–133 (1998)
- 27.31 T. Viklands: Algorithms for the Weighted Orthogonal Procrustes Problem and Other Least Squares Problems, *Ph.D. Thesis* (Umea Univ., Umea 2006)
- 27.32 P.J.G. Teunissen, A. Kleusberg: *GPS for Geodesy*, 2nd edn. (Springer, Berlin 1998)
- 27.33 G. Giorgi: GNSS Carrier Phase-Based Attitude Determination. Estimation and Applications, *Ph.D. Thesis* (Delft Univ. Technology, Delft 2011)
- 27.34 P.J.G. Teunissen: Nonlinear least-squares, *Manuscripta Geodaetica* **15**(3), 137–150 (1990)
- 27.35 P.J.G. Teunissen: The affine constrained GNSS attitude model and its multivariate integer least-squares solution, *J. Geod.* **86**(7), 547–563 (2012)
- 27.36 A.J.V. Dierendonck, P. Fenton, T. Ford: Theory and performance of narrow correlator spacing in a GPS receiver, *Navigation* **39**(3), 265–283 (1992)

- 27.37 A. Simsky, J.M. Sleewaegen, M. Hollreiser, M. Crisci: Performance assessment of Galileo ranging signals transmitted by GSTB-V2 satellites, Proc. ION GNSS, Fort Worth (ION, Virginia 2006) pp. 1547–1559
- 27.38 L.R. Weill: Multipath mitigation using modernized GPS signals: How good can it get?, Proc. ION GPS, Portland (ION, Virginia 2002) pp. 493–505
- 27.39 C.E. Cohen: Attitude Determination Using GPS, Ph.D. Thesis (Stanford Univ., Palo Alto 1992)
- 27.40 C.E. Cohen: Attitude determination. In: *Global Positioning System: Theory and Applications*, Vol. 2, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Reston 1996)
- 27.41 J.L. Crassidis, F.L. Markley, E.G. Lightsey: Global positioning system integer ambiguity resolution without attitude knowledge, J. Guid. Contr. Dyn. **22**(2), 212–218 (1999)
- 27.42 A. Conway, P. Montgomery, S. Rock, R. Cannon, B. Parkinson: A new motion-based algorithm for GPS attitude integer resolution, Navigation **43**(2), 179–190 (1996)
- 27.43 E.G. Lightsey, J.L. Crassidis, F.L. Markley: Fast integer ambiguity resolution for GPS attitude determination, AIAA Guid. Navig. Contr. Conf., Portland (AIAA, Reston 1999) pp. 403–412
- 27.44 Y. Wang, X. Zhan, Y. Zhang: Improved ambiguity function method based on analytical resolution of GPS attitude determination, Meas. Sci. Technol. **18**(9), 2985–2990 (2007)
- 27.45 M.L. Psiaki: Batch algorithm for global-positioning-system attitude determination and integer ambiguity resolution, J. Guid. Contr. Dyn. **29**(1), 1070–1079 (2006)
- 27.46 C.C. Counselman, S.A. Gourevitch: Miniature interferometer terminals for Earth surveying: Ambiguity and multipath with the Global Positioning System, IEEE Trans. Geosci. Remote Sens. **GE-19**(4), 244–252 (1981)
- 27.47 A. Caporali: Basic direction sensing with GPS, GPS World **12**(3), 44–50 (2001)
- 27.48 H.J. Euler, C. Hill: Attitude determination: Exploiting all information for optimal ambiguity resolution, Proc. ION GPS, Palm Springs (ION, Virginia 1995) pp. 1751–1757
- 27.49 J.C. Juang, G.S. Huang: Development of GPS-based attitude determination algorithms, IEEE Trans. Aerosp. Electron. Syst. **33**(3), 968–976 (1997)
- 27.50 Y. Li, K. Zhang, C. Roberts, M. Murata: On-the-fly GPS-based attitude determination using single- and double-differenced carrier phase measurements, GPS Solutions **8**(2), 93–102 (2004)
- 27.51 L.V. Kuylen, P. Nemry, F. Boon, A. Simsky, J.F.M. Lorga: Comparison of attitude performance for multi-antenna receivers, Eur. J. Navig. **4**(2), 1–9 (2006)
- 27.52 R. Monikes, J. Wendel, G.F. Trommer: A modified LAMBDA method for ambiguity resolution in the presence of position domain constraints, Proc. ION GNSS, Long Beach (ION, Virginia 2005) pp. 81–87
- 27.53 A. Hauschild, G. Grillmayer, O. Montenbruck, M. Markgraf, P. Vörsmann: GPS attitude determination for the flying laptop satellite. In: *Small Satellites for Earth Observation*, ed. by R. Sandau, H.P. Röser, A. Valenzuela (Springer, Netherlands 2008)
- 27.54 B. Wang, L. Miao, S. Wang, J. Shen: A constrained LAMBDA method for GPS attitude determination, GPS Solutions **13**(2), 97–107 (2009)
- 27.55 R. Hatch: Instantaneous ambiguity resolution, Proc. Int. Symp. Kinemat. Syst. Geod. Surv. Remote Sens. (KIS), Banff, ed. by K.-P. Schwarz, G. Lachapelle (Springer, New York 1991) pp. 299–308
- 27.56 R.A. Brown: Instantaneous GPS attitude determination, Proc. IEEE PLANS, Monterey (IEEE, Cleveland 1992) pp. 113–120
- 27.57 C. Park, I. Kim, J.G. Lee, G.I. Jee: Efficient ambiguity resolution using constraint equation, Proc. IEEE PLANS, Atlanta (IEEE, Cleveland 1996) pp. 227–284
- 27.58 M.S. Hodgart, S. Purivigraipong: New approach to resolving instantaneous integer ambiguity resolution for spacecraft attitude determination using GPS signals, Proc. IEEE PLANS, San Diego (IEEE, Cleveland 2000) pp. 132–139
- 27.59 P.J.G. Teunissen: Least-squares estimation of the integer GPS ambiguities, Invited Lecture, Section IV Theory and Methodology, IAG Gen. Meet., Beijing (IAG, 1993)
- 27.60 P.J.G. Teunissen: The Least-squares ambiguity decorrelation adjustment: A method for fast GPS integer ambiguity estimation, J. Geod. **70**(1/2), 65–82 (1995)
- 27.61 P. De Jonge, C.C.J.M. Tiberius: The LAMBDA method for integer ambiguity estimation: Implementation aspects, Publ. Delft Comput. Cent. LGR-Series **12**, 1–47 (1996)
- 27.62 P.J.G. Teunissen, P.J. de Jonge, C.C.J.M. Tiberius: Performance of the LAMBDA method for fast GPS ambiguity resolution, J. Navig. **44**(3), 373–383 (1997)
- 27.63 P.J.G. Teunissen: An optimality property of the integer least-squares estimator, J. Geod. **73**(11), 587–593 (1999)
- 27.64 S. Verhagen, P.J.G. Teunissen: New global navigation satellite system ambiguity resolution method compared to existing approaches, J. Guid. Contr. Dyn. **29**(4), 981–991 (2006)
- 27.65 A. Hauschild, O. Montenbruck: GPS-based attitude determination for microsatellites, Proc. ION GNSS, Fort Worth (ION, Virginia 2007) pp. 2424–2434
- 27.66 L. Dai, K.V. Ling, N. Nagarajan: Real-time attitude determination for microsatellite by LAMBDA method combined with Kalman filtering, 22nd AIAA Int. Commun. Satell. Syst. Conf. Exhib. (IC-SSC), Monterey (AIAA, Reston 2004) pp. 1–8
- 27.67 P.J.G. Teunissen, G. Giorgi, P.J. Buist: Testing of a new single-frequency GNSS carrier-phase compass method: Land, ship and aircraft experiments, GPS Solutions **15**(1), 15–28 (2010)

- 27.68 P.J.G. Teunissen: Integer least-squares theory for the GNSS compass, *J. Geod.* **84**(7), 433–447 (2010)
- 27.69 G. Giorgi, P.J.G. Teunissen, S. Verhagen, P.J. Buist: Instantaneous ambiguity resolution in global-navigation-satellite-system-based attitude determination applications: A multivariate constrained approach, *J. Guid. Contr. Dyn.* **35**(1), 51–67 (2012)
- 27.70 G. Giorgi, P.J.G. Teunissen, P.J. Buist: A Search and shrink approach for the baseline constrained LAMBDA: Experimental results, *Int. Symp. GPS/GNSS*, Tokyo, ed. by A. Yasuda (Tokyo Univ. of Marine Science and Technology, Tokyo 2008) pp. 797–806
- 27.71 N. Nadarajah, P.J.G. Teunissen, G. Giorgi: GNSS attitude determination for remote sensing: On the bounding of the multivariate ambiguity objective function. In: *Earth on the Edge: Science for a Sustainable Planet*, ed. by C. Rizos, C. Willis (Springer, Berlin 2014) pp. 503–509
- 27.72 M. Ueno: GPS Attitude for a Berthing Guidance System, Ph.D. Thesis (Universite Laval, Quebec 1999)
- 27.73 G. Giorgi, P.J.G. Teunissen: Low-complexity instantaneous ambiguity resolution with the affine-constrained GNSS attitude model, *IEEE Trans. Aerosp. Electron. Syst.* **49**(3), 1745–1759 (2013)
- 27.74 P.J.G. Teunissen: The probability distribution of the GPS baseline for a class of integer ambiguity estimators, *J. Geod.* **73**(5), 275–284 (1999)
- 27.75 K. Ferguson, J. Kosmalska, M. Kuhl, J.M. Eichner, K. Kepski, R. Abtahi: Three-dimensional attitude determination with the ashtech 3DF 24-channel GPS measurement system, *Proc. ION NTM*, Phoenix (ION, Virginia 1991) pp. 35–41
- 27.76 C.E. Cohen, E.G. Lightsey, B.W. Parkinson: Space flight tests of attitude determination using GPS, *Int. J. Satell. Commun.* **12**(5), 427–433 (1994)
- 27.77 H.J. Kramer: *Observation of the Earth and Its Environment: Survey of Missions and Sensors*, 4th edn. (Springer, Berlin, Heidelberg 2001) pp. 145–156
- 27.78 F.L. Knight: The space test program APEX mission – Flight results, *AIAA/USU Conf. Small Satell.*, Logan (Utah State Univ., Logan 1996) pp. 1–15
- 27.79 D. Freesland, K. Reiss, D. Young, J. Cooper, C.A. Adams: GPS based attitude determination: The REX II flight experience, *AIAA/USU Conf. Small Satell.*, Logan (Utah State Univ., Logan 1996) pp. 1–9
- 27.80 M. Unwin, S. Purivigraipong, A. da Silva Curiel, M. Sweeting: Stand-alone spacecraft attitude determination using real flight GPS data from UOSAT-12, *Acta Astronaut.* **51**(1), 261–268 (2002)
- 27.81 J.C. Adams: Robust GPS Attitude Determination for Spacecraft, Ph.D. Thesis (Stanford Univ., Palo Alto 1999)
- 27.82 H. Uematsu, L. Ward, B.W. Parkinson: Use of global positioning system for gravity probe B relativity experiment and co-experiments, *Adv. Space Res.* **26**(6), 1199–1203 (2000)
- 27.83 S. Gomez: Three years of global positioning system experience on international space station, NASA/TP-2006-213168 (NASA Johnson Space Center, Houston 2006)
- 27.84 M.D. DiPrinzi, R.H. Tolson: *Evaluation of GPS Position and Attitude Determination for Automated Rendezvous and Docking Missions* (NASA, Langley Research Center, Hampton 1994)
- 27.85 J.L. Goodman: *GPS Lessons Learned from the International Space Station, Space Shuttle and X-38*, NASA-CR-2005-213693 (NASA Johnson Space Center, Houston 2005)
- 27.86 C.E. Cohen: Flight tests of attitude determination using GPS compared against an inertial navigation unit, *Proc. ION NTM*, San Francisco (ION, Virginia 1993) pp. 579–587
- 27.87 K.P. Schwarz: Aircraft position and attitude determination by GPS and INS generalized solution of the orthogonal procrustes problem, *Int. Arch. Photogramm. Remote Sens.* **31**(B6), 67–73 (1996)
- 27.88 D. Gebre-Egziabher, R.C. Hayward, J.D. Powell: A low-cost GPS/inertial attitude heading reference system (AHRS) for general aviation applications, *Proc. IEEE PLANS*, Palm Springs (IEEE, Cleveland 1998) pp. 518–525
- 27.89 F. Boon, B.A.C. Ambrosius: Results of real-time applications of the LAMBDA method in GPS based aircraft landings, *Proc. Int. Symp. Kinemat. Syst. Geod. Geomat. Navig. (KIS)*, Banff (Univ. Calgary, Calgary 1997) pp. 339–345
- 27.90 M.J. Moore, C. Rizos, J. Wang, G. Boyd, K. Matthew: A GPS based attitude determination system for an UAV aided by low grade angular rate gyros, *Proc. ION GNSS*, Portland (ION, Virginia 2003) pp. 2417–2424
- 27.91 S. Corbett: GPS for attitude determination and positioning in airborne remote sensing, *Proc. ION GPS*, Salt Lake City (ION, Virginia 1993) pp. 789–796
- 27.92 B.A. Alberts, B.C. Gunter, A. Muis, Q.P. Chu, G. Giorgi, P.J. Buist, C.C.J.M. Tiberius, H. Lindenburg: Correcting strapdown GPS/INS gravimetry estimates with GPS attitude data, *Int. Assoc. Geod. Symp. Grav. Geoid Earth Obs.* **135**, 93–100 (2010)
- 27.93 J.A. Mercer, R.R. Ryan, H.A. Kolve: United States Navy applications of a GPS attitude and position measurement system, *Proc. ION NTM*, Albuquerque (ION, Virginia 1992) pp. 783–791
- 27.94 G. Lachapelle, M.E. Cannon, B. Loncarevic: Shipborne GPS attitude determination during MMST-93, *IEEE J. Ocean. Eng.* **21**(1), 100–105 (1996)
- 27.95 J.A. Kawahara, M. Meakin: Using a GPS antenna array to provide ship heading for a precise integrated navigation system, *Can. Hydrogr. Conf.*, Halifax (Canadian Hydrographic Service, Halifax 1996) pp. 63–69
- 27.96 G. Lu: Development of a GPS Multi-Antenna System for Attitude Determination, Ph.D. Thesis (Univ. Calgary, Calgary 1995)
- 27.97 G. Schleppe: Development of a Real-Time Attitude System Using a Quaternion Parameterization and Non-Dedicated GPS Receivers, Ph.D. Thesis (Univ.

- Calgary, Calgary 1996)
- 27.98 G. Giorgi, P.J.G. Teunissen, T. Gourlay: Instantaneous global navigation satellite system (GNSS)-based attitude determination for maritime applications, *IEEE J. Ocean. Eng.* **37**(3), 348–362 (2012)
- 27.99 T.P. Gourlay, K. Klaka: Full-scale measurements of containership sinkage, trim and roll, *Aust. Nav. Archit.* **11**(2), 30–36 (2007)
- 27.100 M. Ueno, R. Santerre: GPS attitude for a berthing guidance system, *Canad. Aeronaut. Space J.* **45**(3), 264–269 (1999)
- 27.101 Y. Yang, J.A. Farrell: Two antennas GPS-aided INS for attitude determination, *IEEE Trans. Contr. Syst. Technol.* **11**(6), 905–918 (2003)
- 27.102 D.S. De Lorenzo, S. Alban, J. Gautier, P. Enge, D. Akos: GPS attitude determination for a JPALS testbed: Integer initialization and testing, *Proc. IEEE PLANS, Monterey* (IEEE, Cleveland 2004) pp. 762–770
- 27.103 M. O'Connor, T. Bell, G. Elkaim, B.W. Parkinson: Automatic steering of farm vehicles using GPS, 3rd Int. Conf. Precis. Agric., Minneapolis, ed. by P.C. Robert, R.H. Rust, W.E. Larson (American Society of Agronomy, Madison 1996) pp. 767–778
- 27.104 S. Panzieri, F. Pascucci, G. Ulivi: An outdoor navigation system using GPS and inertial platform, *IEEE/ASME Trans. Mechatron.* **7**(2), 134–142 (2002)
- 27.105 J. Borenstein, H.R. Everett, L. Feng: *Where am I? Sensors and Methods for Mobile Robot Positioning* (Univ. Michigan, Ann Arbor 1996)
- 27.106 K.T. Mueller, R. Bortins: GPS locomotive location system for high speed rail applications, *Proc. Int. Symp. Kinemat. Syst. Geod. Geomat. Navig. (KIS)*, Banff (Univ. Calgary, Calgary 2001) pp. 42–51
- 27.107 K.P. Schwarz, M.A. Chapman, M.W. Cannon, P. Gong: An integrated INS/GPS approach to the georeferencing of remotely sensed data, *Photogramm. Eng. Remote Sens.* **59**(11), 1667–1674 (1993)
- 27.108 S. Kocaman: *GPS and INS Integration with Kalman Filtering for Direct Georeferencing of Airborne Imagery*, Geodetic Seminar Report (Institute of Geodesy and Photogrammetry, ETH Henggerberg, Zürich 2003)
- 27.109 S. Knedlik, E. Edwan, J. Zhou, Z. Dai, P. Uboldosold, O. Loffeld: GPS/INS integration for footprint chasing in bistatic SAR experiments, *IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Boston (IEEE, Boston 2008) pp. 459–462
- 27.110 G. Giorgi, P.J.G. Teunissen, S. Verhagen, P.J. Buist: Testing a new multivariate GNSS carrier phase attitude determination method for remote sensing platforms, *Adv. Space Res.* **46**(2), 118–129 (2010)
- 27.111 R. Anderson, H.R. Bilger, G.E. Stedman: Sagnac effect: A century of Earth-rotated interferometers, *Am. J. Phys.* **62**(11), 975–985 (1994)
- 27.112 J.D. Fairbank, P.F. Michelson, C.W. Everitt: *Near Zero: New Frontiers of Physics* (W.H. Freeman and Company, New York 1988)
- 27.113 J. Farrell, B. Matthew: *The Global Positioning System and Inertial Navigation* (McGraw-Hill, New York 1999)
- 27.114 S. Alban: Design and Performance of a Robust GPS/INS Attitude System for Automobile Applications, Ph.D. Thesis (Stanford Univ., Palo Alto 2004)
- 27.115 M. Brenner: Integrated GPS/inertial fault detection availability, *Navigation* **43**(2), 339–358 (1996)
- 27.116 C. Kreye, B. Eissfeller, D. Sanroma, T. Lück: Performance analysis and development of a tightly coupled GNSS/INS system, *Proc. 9th St. Petersburg Int. Conf. Integr. Navig. Syst.*, St. Petersburg (Elektropribor, St. Petersburg 2002)
- 27.117 A.B. Chatfield: *Fundamentals of High Accuracy Inertial Navigation*, Progress in Astronautics and Aeronautics, Vol. 174 (AIAA, Reston 1996)
- 27.118 D.T. Knight: Achieving modularity with tightly-coupled GPS/INS, *Proc. IEEE PLANS, Monterey* (IEEE, Cleveland 1992) pp. 426–432
- 27.119 P.D. Groves, C.J. Mather: Receiver interface requirements for deep INS/GNSS integration and vector tracking, *J. Navig.* **63**(3), 471–489 (2010)
- 27.120 P.C. Hughes: *Spacecraft Attitude Dynamics*, 1st edn. (Dover Publications, Mineola 1997)
- 27.121 M.J. Sidi: *Spacecraft Dynamics and Control*, 1st edn. (Cambridge Univ. Press, Cambridge 1997)

GNSS/INS Inte

28. GNSS/INS Integration

Jay A. Farrell, Jan Wendel

This chapter discusses the role of global navigation satellite systems (GNSSs) and inertial measurements in the estimation of the state vector for a maneuvering system. The chapter considers the main objectives of accuracy, continuity, availability, and integrity; and, the contributions that the different types of sensors make toward achieving these objectives. The chapter includes an example design. Then, the chapter reviews the concepts of loose, tight, and ultratight or deeply coupled systems. Throughout, the advantages, disadvantages, and tradeoffs between alternative approaches are discussed.

28.1	State Estimation Objectives	812	28.4	Strapdown Inertial Navigation	818
28.2	Inertial Navigation	813	28.4.1	Coordinate Systems	818
28.2.1	Problem Statement	813	28.4.2	Attitude Calculations	819
28.2.2	Sensor Models	813	28.4.3	Velocity Calculations	821
28.2.3	INS Computations	813	28.4.4	Position Calculations	821
28.2.4	INS Error State	814	28.5	Analysis of Error Effects	822
28.2.5	Performance Characterization	814	28.5.1	Short-Term Effects	822
28.3	Inertial Sensors	815	28.5.2	Long-Term Effects	823
28.3.1	Gyroscopes	815	28.6	Aided Navigation	824
28.3.2	Accelerometers	816	28.7	State Estimation	824
28.3.3	Inertial Sensor Errors	816	28.8	GNSS and Aided INS	825
			28.8.1	Loose (Position Domain) Coupling	825
			28.8.2	Tight (Observable Domain) Coupling	826
			28.8.3	Ultra-Tight or Deep Coupling	826
			28.8.4	Illustrative Comparison	828
			28.9	Detailed Example	828
			28.9.1	System Model	828
			28.9.2	Measurement Models	831
			28.10	Alternative Estimation Methods	835
			28.10.1	Standalone GNSS	835
			28.10.2	Advanced Bayesian Estimation	837
			28.11	Looking Forward	838
			References		839

Although global navigation satellite systems (GNSSs) are often referred to as positioning systems, when used in combination with inertial sensors, they have a much greater utility in helping to maintain the accuracy of the system state, which includes position, velocity, acceleration, attitude, and angular rate [28.1, Sect. 2.4]. These quantities are necessary in applications that in-

volve safety augmentation, control, or trajectory or mission planning. In addition, a navigation system that maintains the system state can have improved performance, perhaps coasting through short durations while GNSS signals are not available. These concepts are discussed in greater detail in the following subsections.

28.1 State Estimation Objectives

For commercial applications, cost is a critical design metric. The fact that the cost of computation and sensors has been rapidly declining over the last few decades has given rise to a large and growing interest in aided inertial navigation systems for commercial applications. In previous decades, such systems were only cost effective for military and life-critical commercial vehicles. Now they are becoming commercially available in hand-held devices.

While navigation system accuracy metrics (e.g., horizontal position error less than 2.0 m with 95% probability) are widely understood, other navigation system metrics such as integrity, continuity, and availability require a little more explanation [28.2]. These terms are also introduced in Sect. 12.1.1:

- *Integrity* relates to the extent to which the information supplied by the navigation system can be trusted or to the ability of the navigation system to detect and to provide timely warning to the user about when the specified accuracy should not be trusted.
- *Continuity* relates to the probability that a specified level of accuracy will be maintained throughout a given operation or experiment, assuming that the specification is met at initialization.
- *Availability* relates to the percentage of time that a specified level of accuracy, integrity, and continuity is available and useable within a specified area.

The navigation system design tradeoffs to address the above metrics are diverse; however, the following are somewhat fundamental. Addition of sensors typically enhances accuracy, integrity, and availability, but at an increasing cost and additional risk of sensor failure.

Integrity is enhanced via redundant signals that enable detection and possible removal of anomalous events that might otherwise affect the navigation system accuracy in an undetected manner. Signal redundancy can be achieved directly. Examples include using signals from more than four Global Positioning System (GPS) satellites in a position solution or using signals from more than one GNSS [28.3, 4]. Signal redundancy can also be analytic or indirect [28.5, 6]. This chapter focus on the indirect form of redundancy that is achieved by combining GNSS with specific force and angular rate measurements provided by an inertial measurement unit (IMU). The IMU measurements are mathematically manipulated to maintain high-bandwidth, high-rate estimates of the IMU position, velocity, acceleration, attitude, and angular rate.

These estimates allow the GNSS measurements to be predicted. The fact that the GNSS predicted and actual measurements can be compared shows that the IMU and GNSS systems are analytically or indirectly redundant.

The utility of redundant measurements is enhanced when the factors that affect measurement availability are distinct for the different measurements. Reception of GNSS signals can be affected by radio frequency interference, terrain, foliage, and man-made structures. When a sufficient number of GNSS satellite signals are unavailable, then no GNSS position solution can be computed. The IMU measurements are not affected by terrain, foliage, external electromagnetic signals, or man-made structures. The fact that the IMU measurements are immune to the factors that disrupt GNSS signals allows the combined system to *coast through* periods of GNSS outages, enhancing continuity and availability.

All sensors are affected by measurement noise and (possibly time-varying) calibration errors. The inertial navigation system (INS) that computes the estimated vehicle state from the IMU measurements is an integrative process. The integration of measurement noise and calibration errors causes the INS estimate of the vehicle state $\hat{\mathbf{x}}(t)$ to diverge from the true vehicle state $\mathbf{x}(t)$ as time passes. This divergence can be accurately modeled as a function of time. High-end INSs are designed to meet a specified accuracy for a stated duration of time (months). Commercial microelectromechanical system (MEMS)-based INS typically require aiding by external sensors at much higher rates (seconds to minutes); otherwise, the error would eventually become too large to be useful. In a well-designed system, aiding will both correct the estimated state and improve the system calibration such that the rate of divergence will be slower in the future. The GNSS/IMU integrated navigation system accuracy as a function of the sensor characteristics and duration of outage can be accurately characterized.

To design an aided INS system well, the analyst must understand the INS kinematic model, the GNSS and IMU sensor models, the tradeoffs between the various integration approaches, methods for estimating the state and sensor calibration errors, factors affecting the observability of those errors, methods for detecting invalid measurements, and methods for analyzing system performance. The purpose of this chapter is to highlight various factors that are of critical importance and to motivate further investigation using any of the various books that discuss these topics in greater detail [28.7–12].

28.2 Inertial Navigation

This section introduces key concepts related to inertial navigation. The discussion has been organized to separate models and theory (i. e., Sects. 28.2.1, 28.2.2, and 28.2.4) from items that are available on and computed by the navigation computer (i. e., Sects. 28.2.3 and 28.2.5).

28.2.1 Problem Statement

The vehicle carrying the INS will be referred to as the *rover*. The rover navigation state vector is $\mathbf{x} = [\mathbf{p}^\top, \mathbf{v}^\top, \mathbf{q}^\top]^\top \in \mathbb{R}^n$, where \mathbf{p} is the position, \mathbf{v} is the velocity, and \mathbf{q} is a representation of the attitude of the rover with respect to the navigation frame of reference [28.13]. Attitude parameterization is discussed further in Sect. 28.4.2. The inertial acceleration and angular rate vector is $\mathbf{u} = [\mathbf{a}^\top, \boldsymbol{\omega}^\top]^\top \in \mathbb{R}^6$. The kinematic equation for the rover navigation state vector is

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}), \quad (28.1)$$

where the vector $\mathbf{f}: \mathbb{R}^n \times \mathbb{R}^6 \rightarrow \mathbb{R}^n$ is accurately known (Sect. 28.4 or [28.8, Chap. 11]). Note that both \mathbf{x} and \mathbf{u} are unknown.

The navigation system assumes the availability of IMU measurements $\tilde{\mathbf{u}}_k = \tilde{\mathbf{u}}(\tau_k)$ where $\tau_k = k\tau$ for $k = 0, 1, 2, \dots$, aiding measurements $\tilde{\mathbf{y}}(t_j)$ where $t_j = jT$ with $\tau \ll T$, and an initial distribution for the state $\mathbf{x}(0) \sim \mathcal{N}(\mathbf{x}_0, \mathbf{P}_{x_0})$. The notation $\mathbf{x}(0) \sim \mathcal{N}(\mathbf{x}_0, \mathbf{P}_{x_0})$ means that $\mathbf{x}(0)$ is normally distributed with mean \mathbf{x}_0 , and covariance matrix \mathbf{P}_{x_0} . This notation will be used throughout this chapter.

The navigation system design problem is to use the IMU and aiding sensor information, the initial condition distribution, and (28.1) to compute an estimate $\hat{\mathbf{x}}(t)$ for all $t = k\tau$ for $k = 0, 1, 2, \dots$. The ideal design would be optimal within the constraints imposed by the system cost.

28.2.2 Sensor Models

The IMU measurements $\tilde{\mathbf{u}}$ are related to the inertial quantities in \mathbf{u} through a set of calibration factors denoted by $\mathbf{c}_u(t) \in \mathbb{R}^{n_u}$ which could, for example, account for the scale factor (SF), alignment, or bias errors. An example of a simple IMU measurement model is

$$\tilde{\mathbf{u}}(t) = \mathbf{u}(t) + \mathbf{c}_u(t) + \mathbf{v}_u(t), \quad (28.2)$$

$$\dot{\mathbf{c}}_u(t) = -\lambda_u \mathbf{c}_u(t) + \mathbf{v}_{c_u}(t). \quad (28.3)$$

The IMU provides only the measurement $\tilde{\mathbf{u}}(t)$. The quantities \mathbf{c}_u , $\lambda_u > 0$, $\mathbf{v}_u(t)$, and $\mathbf{v}_{c_u}(t)$ are all unknown. Typical assumptions are that $\mathbf{v}_u(t)$ and $\mathbf{v}_{c_u}(t)$

are both white Gaussian noise processes with power spectral density (PSD) [28.14–16] matrices denoted by \mathbf{Q}_1 and \mathbf{Q}_2 , respectively. The IMU manufacturer typically provides *Allan* variance [28.17] information to the designer, which is useful for specifying parameters such as λ_u , \mathbf{Q}_1 , and \mathbf{Q}_2 to quantify the IMU state space error model.

Each vector of simultaneous aiding measurements is modeled as

$$\tilde{\mathbf{y}}(t_j) = \mathbf{h}(\mathbf{x}(t_j), \mathbf{c}_y(t_j)) + \boldsymbol{\eta}_y(t_j), \quad (28.4)$$

$$\dot{\mathbf{c}}_y(t) = -\lambda_y \mathbf{c}_y(t) + \mathbf{v}_{c_y}(t), \quad (28.5)$$

where $\mathbf{c}_y(t) \in \mathbb{R}^{n_y}$ is a vector of sensor calibration variables. The aiding sensor provides only the measurement $\tilde{\mathbf{y}}(t)$. The quantities $\mathbf{c}_y(t)$, $\lambda_y > 0$, $\boldsymbol{\eta}_y(t)$, and $\mathbf{v}_{c_y}(t)$ are all unknown. The noise process $\boldsymbol{\eta}_y(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}(t))$ is assumed to be white with covariance matrix $\mathbf{R}(t)$. The noise process $\mathbf{v}_{c_y}(t)$ is assumed to be white and Gaussian with the PSD matrix denoted by \mathbf{Q}_3 . In the context of GNSS, \mathbf{c}_y could be used to account for the time correlated multipath errors, for example.

Stochastic processes and modeling are important. Topics such as white noise, Markov processes, and Allan variance are discussed in [28.14, 15].

28.2.3 INS Computations

Given the initial condition for the estimated rover state vector $\hat{\mathbf{x}}(0) = \mathbf{x}_0$ and the IMU measurements $\tilde{\mathbf{u}}$, the INS computes an estimate of the rover state by solving

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{f}(\hat{\mathbf{x}}(t), \hat{\mathbf{u}}(t)) \quad (28.6)$$

via numeric integration in real time. The INS only has IMU and aiding measurements at discrete time instants; therefore, the INS numerically solves

$$\begin{aligned} \hat{\mathbf{x}}(\tau_{k+1}) &= \boldsymbol{\phi}(\hat{\mathbf{x}}(\tau_k), \hat{\mathbf{u}}(\tau_k)) \\ &= \hat{\mathbf{x}}(\tau_k) + \int_{\tau_k}^{\tau_{k+1}} \mathbf{f}(\hat{\mathbf{x}}(\tau), \hat{\mathbf{u}}(\tau)) d\tau. \end{aligned} \quad (28.7)$$

The result of the numeric integration of (28.7) is the INS state estimate $\hat{\mathbf{x}}(\tau_{k+1})$ given $\hat{\mathbf{x}}(\tau_k)$ and $\hat{\mathbf{u}}(\tau_k)$.

The quantity $\hat{\mathbf{u}}(t)$ is computed from $\tilde{\mathbf{u}}(t)$ using the best available estimates $\hat{\mathbf{c}}_u(t)$ of the IMU calibration factors. For example, given the IMU model of (28.2), $\hat{\mathbf{u}}(t) = \tilde{\mathbf{u}}(t) - \hat{\mathbf{c}}_u(t)$.

A variety of INS mechanization design factors influence the system performance, such as choice of reference frames, IMU model, numeric integration method, and coning and sculling computations [28.18].

Note that as long as the IMU continues to provide measurements, the INS will continue to use those measurements to provide state estimates. The ability of the INS to compute the state estimate is independent of external magnetic fields, terrain, and so on. This enhances continuity, integrity, and availability with respect to GNSS only solutions. It must be noted, however, that the accuracy of the unaided INS solution will deteriorate as the time from the last aiding signal increases.

Also, note that the bandwidth of the INS state estimation system is determined by the bandwidth of the IMU, which can be in the 0.1–1.0 kHz range. This is critical when the state estimates are inputs to a control system.

The numeric integration repeats using the sequence of IMU measurements to propagate the state measurements between the times of aiding measurements. Given the INS state vector $\hat{\mathbf{x}}$, the estimated sensor calibration parameter vector $\hat{\mathbf{c}}_y$, and the model of the aiding measurements in (28.4), the predicted value of the aiding measurements is

$$\hat{\mathbf{y}}(t_j) = \mathbf{h}(\hat{\mathbf{x}}(t_j), \hat{\mathbf{c}}_y(t_j)) . \quad (28.8)$$

The aiding measurement times can be unequally spaced in time without causing any complications. Missing measurements are also easy to accommodate.

Let

$$\hat{\mathbf{z}} = [\hat{\mathbf{x}}^\top, \hat{\mathbf{c}}_u^\top, \hat{\mathbf{c}}_y^\top]^\top \in \mathbb{R}^{n+n_y+n_u} .$$

The vector $\hat{\mathbf{z}}$ contains all the quantities required by the INS to compute (28.7) and (28.8). The on-line navigation algorithms are designed to compute $\hat{\mathbf{z}}(t)$ as an estimate of $\mathbf{z}(t) = [\mathbf{x}^\top, \mathbf{c}_u^\top, \mathbf{c}_y^\top]^\top$. Although $\hat{\mathbf{x}}$ is the only quantity required for the navigation solution, estimation of $\hat{\mathbf{c}}_u$ enhances the accuracy of the time propagation step of (28.7), and estimation of $\hat{\mathbf{c}}_y$ enhances the accuracy of the aiding measurement correction step to be discussed in Sect. 28.7.

28.2.4 INS Error State

Due to initial condition errors, calibration errors, and measurement noise, a state estimation error

$$\delta \mathbf{z}(t) = \mathbf{z}(t) - \hat{\mathbf{z}}(t) \quad (28.9)$$

develops over time. The state error vector evolves in time according to the standard (linearized) discrete-time model

$$\delta \mathbf{z}(t_{j+1}) = \Phi_j \delta \mathbf{z}(t_j) + \mathbf{v}_{\delta j} , \quad (28.10)$$

where $\mathbf{v}_{\delta j}$ is the discrete-time noise resulting from the accumulation of the random process vector

$$\mathbf{v}^\top(t) = [\mathbf{v}_u^\top(t), \mathbf{v}_{c_u}^\top(t), \mathbf{v}_{c_y}^\top(t)]$$

by the system over the time interval $t \in [t_j, t_{j+1}]$. The error state transition matrix Φ_j , the process noise $\mathbf{v}_{\delta j} \sim N(\mathbf{0}, \mathbf{Q}_j)$, and the stochastic properties of this state estimation error are well understood [28.7–9, 11, 12]. The on-line computations provide both \mathbf{Q}_j and Φ_j [28.8, Sect. 7.2.5.2].

Note that the estimation problem has three quantities changing as functions of time: $\mathbf{z}(t)$ is unknown, $\hat{\mathbf{z}}(t)$ is available on the navigation computer, and $\delta \mathbf{z}(t)$ is unknown. Of these three quantities, only two of the vectors are linearly independent. Given any two of the vectors, the third vector can be computed by manipulation of (28.9).

As previously discussed, the INS is an integrative process. Integration decreases the effects of high-frequency sensor errors (e.g., high-frequency content in \mathbf{v}_{c_u} and \mathbf{v}_u), but amplifies the effects of low frequency sensor errors (e.g., \mathbf{c}_u). To gain market share, IMU manufacturers have incentives to remove all deterministic error effects and to decrease λ_u , \mathbf{c}_u , \mathbf{Q}_1 , and \mathbf{Q}_2 . Therefore, $\delta \mathbf{z}(t)$ is a slowly varying random process that will be estimated in real time.

28.2.5 Performance Characterization

Given the error covariance $\mathbf{P}_{z_j} = \text{cov}(\delta \mathbf{z}(t_j))$ valid at time t_j , using (28.10) the error covariance at time t_{j+1} can be predicted as

$$\mathbf{P}_{z_{j+1}} = \Phi_j \mathbf{P}_{z_j} \Phi_j^\top + \mathbf{Q}_j . \quad (28.11)$$

This equation models the time propagation of INS errors. The effect of measurement corrections will be discussed in Sect. 28.7. Equation (28.11) can be iterated as necessary to compute \mathbf{P}_{z_i} for $t_i > t_j$. Without aiding, the diagonal of \mathbf{P}_{z_i} typically increases with time. The fact that \mathbf{P}_{z_i} is available in real time facilitates integrity analysis, as the system can predict the time duration until specified accuracy limits are expected to be violated.

Such performance predictions are dependent on past motion and assumptions regarding future motion. Given the same satellite configuration, different prior motion patterns will affect the system observability, and hence, determine the current value of \mathbf{P}_{z_j} . Assumptions about the future motion determine the extent to which each portion of \mathbf{P}_{z_j} affects \mathbf{P}_{z_i} for $t_i > t_j$. These topics relate to the subject of *observability* [28.19–22].

Note that in the GNSS-aided INS applications, the portions of Φ_j and \mathbf{Q}_j attributable to the IMU and INS are not *tuning parameters*. They are physical quantities defined by the kinematics of the system and the IMU characteristics.

28.3 Inertial Sensors

The purpose of this section is to introduce the different categories of inertial sensors, to discuss their error characteristics, and to give specific examples of the inertial sensor calibration quantities $c_u(t)$ as defined in Sect. 28.2.2.

A general nomenclature (e.g., \mathbf{a}_{ib}^b) is required to unambiguously identify quantities like accelerations, velocities, angular rates, and so on. In this nomenclature, the upper index indicates the reference frame in which the respective quantity is represented. The second lower index defines which coordinate frame moves with respect to the coordinate frame specified by the first lower index. Therefore, \mathbf{a}_{ib}^b is the acceleration of the body frame with respect to the inertial frame represented in body frame coordinates. The common reference frames are defined in Sect. 28.4.1.

28.3.1 Gyroscopes

A gyroscope provides measurements of its (inertial) angular rate vector defined relative to its sensitive axis, that is, ω_{ib}^b . Alternatively, the integral of the angular rate over the sampling time interval can be provided, which is referred to as an angle increment. Different types of gyroscopes suitable for strapdown INS implementations can be distinguished [28.10, 23–25].

Vibrating Structure Gyroscopes

Vibrating structure gyroscopes (VSGs) measure the angular rate by observing the Coriolis acceleration that occurs when the sensor rotates. An example of a possible realization is a proof mass that is forced to perform a linear oscillatory motion in one direction, namely the drive direction. The sense direction is perpendicular to the drive direction. The sensitive axis is perpendicular to both the sense and the drive directions. When a rotation occurs around the sensitive axis, the Coriolis acceleration causes the proof mass to follow an ellipsoidal trajectory in the plane perpendicular to the sensitive axis. The amplitude of this motion in the sense direction is proportional to the angular rate and is measured capacitively. Various alternative realizations of this principle exist that use a string, a beam, a tuning fork, or other structures as a vibrating element. At present, this type of gyroscope is usually a low-cost device with poor accuracy, mostly implemented using MEMS technology. MEMS technology combines integrated circuit technology with small mechanical structures, typically on a single chip. These mechanical structures, for example, a proof mass, are generated using different techniques such as surface micromachining, etching, and photolithographic processes. An

example for a MEMS gyroscope manufactured in dry etching technology is shown in Fig. 28.1.

Fiber Optic Gyroscopes

Fiber optic gyroscopes (FOGs) take advantage of the Sagnac effect, which is based on the fact that the speed of light is independent of the speed of the light source. Laser light is sent clockwise and counter-clockwise through a fiber optic coil. When the structure rotates while the light is traveling through the coil, the optical path length for one of the beams is increased, whereas it is reduced for the other beam. The resulting phase shift between the beams is proportional to the angular rate and is measured interferometrically. Usually, FOGs offer better accuracy than MEMS gyroscopes.

Ring Laser Gyroscopes

Ring laser gyroscopes (RLGs) are also based on the Sagnac effect; the closed optical path is realized using appropriately placed mirrors. A clockwise and a counter-clockwise laser beam establish standing waves within the structure. If the optical path lengths change due to rotation, the wavelength of one beam increases whereas the wavelength of the other one decreases to maintain the standing waves. This causes a shift in the laser frequencies. One of the mirrors is semi-transparent, which allows the frequency difference of the laser beams to be measured by interfering them outside of the cavity. In case no rotation is present, the interference pattern is stationary, otherwise the interference pattern moves, which is detected using photodiodes. Ring laser gyroscopes offer the best performance of the gyroscopes discussed so far.

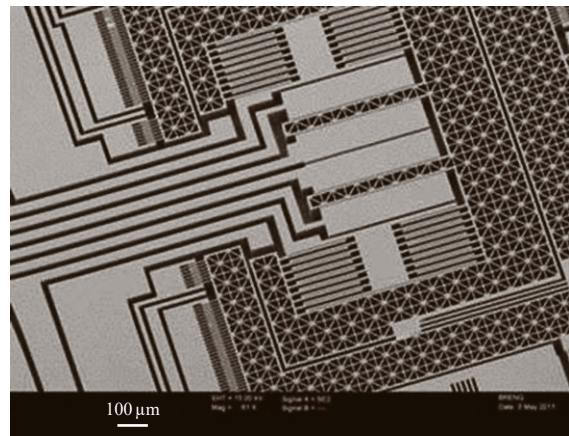


Fig. 28.1 MEMS gyroscope manufactured in dry etching technology (courtesy of Northrop Grumman LITEF GmbH)

Only a simplified view on the most common gyroscopes is provided, other types like the spinning mass gyroscope, or the nuclear magnetic resonance gyroscope, are beyond the scope of this discussion.

28.3.2 Accelerometers

Most practically relevant accelerometers measure the specific force using a proof mass connected to the accelerometer case by a spring. In free space, away from any large masses exerting gravitational effects, for a nonaccelerating accelerometer, the equilibrium position of the proof mass relative to the case would be interpreted as zero acceleration. When the case undergoes an external force, the acceleration is transmitted to the proof mass through the spring deflection. This deflection is measured and scaled to units of acceleration. Near the Earth (or any other large mass), an accelerometer that is not accelerating will have its proof mass deflected from its equilibrium position by the effects of gravity, measuring one g . Conversely, an accelerometer that is in free fall accelerating at one g has its proof mass in its equilibrium position relative to the case, and measures zero acceleration. Both scenarios are illustrated in Fig. 28.2.

Accelerometers measure the specific force vector, f_{ib}^b , which is the difference between the acceleration of an accelerometer triad with respect to an inertial frame, a_{ib}^b , and the local gravity vector, g^b . This can be expressed as

$$f_{ib}^b = a_{ib}^b - g^b. \quad (28.12)$$

Therefore, acceleration is computed from the measured specific force vector by compensating for the effects of gravity.

Two types of accelerometers [28.10, 23–25], the pendulous accelerometer and the vibrating beam accelerometer, provide suitable examples.

Pendulous Accelerometer

Pendulous accelerometers are often manufactured from micromachined silicon by etching. Silicon is removed

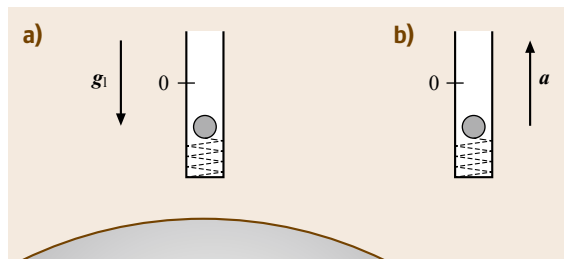


Fig. 28.2a,b Two indistinguishable scenarios. (a) Accelerometer at rest in the Earth's gravity field. (b) Accelerometer accelerated in upward direction, no gravity

from the wafer so that a mass–spring system results. The deflection of the proof mass from equilibrium can be measured capacitively. In open-loop operation, the specific force is calculated from measured deflection of the proof mass. The accuracy of this mode of operation is limited, as the nonlinearity of the small silicon bridge acting as a spring is not known perfectly. In closed-loop operation, a force feedback is used to reset any deflection of the proof mass from equilibrium, often electrostatically. Hereby, the current required to maintain the proof mass in the equilibrium position is a measure of the specific force acting on the proof mass. As an example, the proof mass of a silicon MEMS accelerometer manufactured in wet etching technology is shown in Fig. 28.3.

Vibrating-Beam Accelerometer

Vibrating-beam accelerometers (VBAs) use beams of quartz or silicon that constrain the motion of the proof mass in the direction of the sensitive axis. The specific force acting in this direction compresses one beam and stretches the other. Hereby, the resonant frequencies of the beams change, the resulting frequency difference can be measured with high accuracy, allowing assessment of the specific force acting in the direction of the sensitive axis.

28.3.3 Inertial Sensor Errors

The measurement of the angular rate and specific force vectors using inertial sensors is always prone to imperfections, referred to as sensor errors. A portion of the effect of these errors is removed by factory calibration. When integrated with GNSS, the remaining systematic errors can be estimated by the data fusion algorithm and compensated. However, some imperfections always remain that cause the navigation solution of the INS to deteriorate with time, so that persistent aiding of the INS is mandatory.

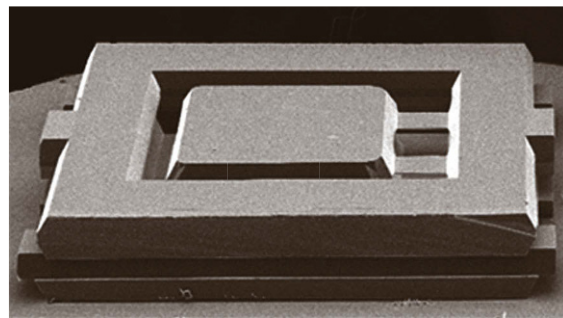


Fig. 28.3 MEMS accelerometer manufactured in wet etching technology (courtesy of Northrop Grumman LITEF GmbH)

More detailed IMU error models than that described in Sect. 28.2.2 are often useful. A specific model for the measurements provided by a triad of inertial sensors at epoch k , may it be gyroscopes or accelerometers, is given by

$$\tilde{\mathbf{u}}_k = \mathbf{M}\mathbf{u}_k + \mathbf{b}_{u,k} + \mathbf{v}_{u,k}, \quad (28.13)$$

where

$$\mathbf{M} = \begin{pmatrix} 1 + s_x & \delta_{z_x} & \delta_{y_x} \\ \delta_{z_y} & 1 + s_y & \delta_{x_y} \\ \delta_{y_z} & \delta_{x_z} & 1 + s_z \end{pmatrix}. \quad (28.14)$$

Here, the components s_x , s_y , and s_z represent the sensor scale factor errors; the off-diagonal elements in this matrix are misalignment errors. The vector $\mathbf{b}_{u,k}$ denotes additive sensor biases, and $\mathbf{v}_{u,k}$ is the sensor noise vector.

In the following, these errors are discussed briefly. Hereby, it has to be kept in mind that each of these errors has a fixed component that is always present, an additional component that varies from run to run, a further component that varies during the operation of the unit, and a component that is temperature dependent. However, depending on the specific sensor, some of these components might not be relevant and can be ignored.

Typically, the designer specifies a Gauss–Markov state space model for these random processes [28.15], using the manufacturer specifications, so that these terms can be estimated (i. e., calibrated) while in use.

Scale Factor

The scale factor error has a linear and a nonlinear component. The linear component causes an error that is proportional to the true specific force or angular rate. For the error caused by the SF nonlinearity, it is often sufficient to assume a quadratic dependency relative to the true value.

Misalignment

Misalignment refers to the nonorthogonality of the sensitive axes of the sensors in a sensor triad. Due to this nonorthogonality, each sensor also senses a part of the specific force or angular rate perpendicular to the nominal direction of the sensitive axis of the sensor.

Bias

Ideally, the bias vector $\mathbf{b}_{u,k}$ is the average reading each sensor shows in case the quantity to be measured, angular rate or specific force, is zero. More practically, the bias term represents a slowly varying random quantity. In addition, especially for MEMS gyroscopes, the bias can have a component that is dependent on the specific force, known as the g -dependent bias.

Noise

The sensor-inherent noise is usually modeled accurately as white, zero-mean, and Gaussian. It is specified in datasheets via its PSD or Allan variance. Most inertial sensors are integrating sensors, for which the relation between PSD R and variance R_k is given by

$$R_k = \frac{R}{\tau}, \quad (28.15)$$

where τ is the IMU sampling period, the reciprocal of the sampling rate. The PSD offers a way to describe the sensor noise independent from the sampling rate. For gyroscopes, the square root of the PSD is referred to as angular random walk (ARW). For accelerometers, this is denoted as velocity random walk (VRW). The PSD allows the designer to assess easily the effect of the ARW and VRW noise effects.

As an example, let us consider the construction of a gyroscope random walk error model. Let μ_k denote white, Gaussian noise with standard deviation $\sigma_\mu = 1$. Let $v_{\omega,k}$ denote the gyroscope noise with variance R_k . Drawing μ_k from a random number generator, the sensor noise can be generated using

$$v_{\omega,k} = \sqrt{R_k} \mu_k = \sqrt{\frac{R}{\tau}} \mu_k. \quad (28.16)$$

Most gyroscopes provide angular increments. For the noise corrupting such an angular increment

$$\Delta\theta_k = \tau v_{\omega,k} = \sqrt{R\tau} \mu_k, \quad (28.17)$$

with variance

$$\sigma_{\Delta\theta}^2 = R\tau \sigma_\mu^2 = R\tau. \quad (28.18)$$

If m angular increments are summed over a time interval $T = m\tau$, the variance of the resulting angle error is given by

$$\sigma_\theta^2 = m \sigma_{\Delta\theta}^2 = m R \tau = R T, \quad (28.19)$$

the standard deviation is given by

$$\sigma_\theta = \sqrt{R} \sqrt{T}. \quad (28.20)$$

Specifying the sensor noise via the square root of the PSD R allows the analyst to calculate the standard deviation of the resulting angle or velocity error simply by multiplying with the square root of the time interval T .

However, the dominant contribution to the noise corrupting the IMU measurements is often not the sen-

Table 28.1 Gyroscope errors

Error	RLG	FOG	VSG
Bias g -indep. ($^{\circ}/h$)	0.001–10	0.1–100	1–3600
g -dep. ($^{\circ}/h/g$)	0	1	10–200
SF (ppm) linear	1–100	100–1000	$< 10^5$
SF (ppm) nonlinear	10^{-5}	$> 10^{-4}$	0.01
ARW ($^{\circ}/\sqrt{h}$)	0.001	0.03–0.1	> 0.06

sor inherent noise, but vibration-induced noise, which is usually non-white [28.26]. Furthermore, in the presence of vibrations, SF nonlinearities and misalignment

Table 28.2 Accelerometer errors

Error	VBA	Pendulous
Bias (mg)	0.1–1	0.1–10
Bias stability (mg)	0.1	1
SF (ppm)	100	1000
VRW ($\frac{m}{s}/\sqrt{h}$)	0.01	0.04

errors cause a sensor error that appears like a vibration-dependent bias, known as a vibration rectification error (VRE) [28.12].

Tables 28.1 and 28.2 present typical orders-of-magnitude of the accelerometer and gyroscope errors.

28.4 Strapdown Inertial Navigation

In a strapdown inertial navigation system, the IMU is rigidly attached to the rover. This is in contrast with gimbaled inertial systems where the IMU is attached to a stabilized platform that maintains its inertial orientation as the rover maneuvers. For gimbaled systems, the rover attitude is measured from the angles of the platform gimbals relative to the rover. Strapdown systems are typically significantly lower in cost and size than gimbaled systems.

In a strapdown algorithm (SDA), the measurements provided by the IMU gyros are processed to maintain an estimate of the rover attitude, so that the accelerometer measurements can be processed to compute the navigation frame velocity and position. The attitude, position, and velocity are each portions of the rover state vector that are propagated forward in time by integrating the IMU measurements through the system kinematic model (Sect. 28.2.3). Without aiding, the deviation of the SDA state from the true rover state increases with time. The speed of growth of these errors depends on the accuracy of the initialization, and on the IMU quality. Table 28.3 lists typical performance and cost for different IMU grades.

The purpose of the following sections is to clearly define the function f and rover state vector x in (28.1) of Sect. 28.2.1, which form the basis for the strapdown calculations.

Table 28.3 IMU grades

IMU grade	Bias (mg)	Bias ($^{\circ}/h$)	Cost (\$)
Marine	0.01	0.001	$\gg 100$ k
Aviation	0.03–0.1	0.01	100 k
Intermed.	0.1–1	0.1	20–50 k
Tactical	1–10	1–100	2–30 k
Consumer	> 10	> 100	≥ 10

28.4.1 Coordinate Systems

The strapdown calculations involve several coordinate systems, which are defined in the following and illustrated in Fig. 28.4.

Inertial Frame

This frame is denoted with the index i , the x - and y -axes lie in the Earth's equatorial plane and are fixed (i. e., not rotating or accelerating). The z -axis coincides with the Earth's rotational axis.

Earth Frame

This frame, referred to as Earth-centered Earth-fixed (ECEF) frame, is denoted with index e . The x - and

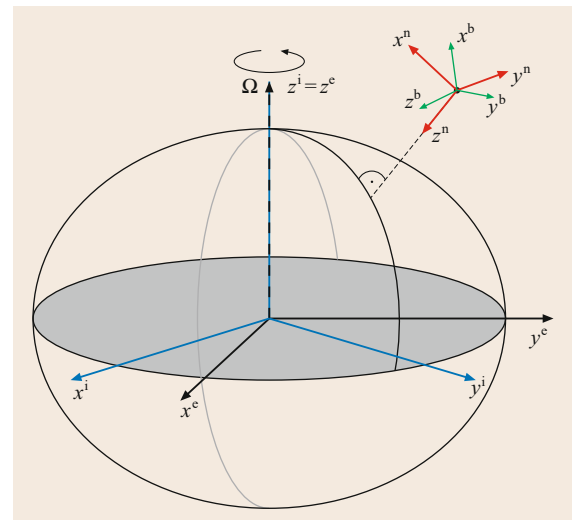


Fig. 28.4 Relative positions and orientations of the various coordinate systems

y -axes lie in the Earth's equatorial plane, the x -axis intersects the Greenwich meridian. The z -axis coincides with the Earth's rotational axis. The e -frame rotates with respect to the i -frame with angular rate ω_{ie} .

Navigation Frame

The origin of this frame is within the vehicle, the axes are aligned with the directions north, east, and local vertical (i.e., down). Therefore, this frame is also called local level or NED frame. This frame is denoted with the index n .

Body Frame

The origin of this frame coincides with the origin of the navigation frame, which is often defined to be the location of the IMU. The axis of the body frame is usually aligned with the directions front, right, and down as viewed from the vehicle. For the sake of simplicity, it is assumed in the following that these axes also coincide with the sensitive axes of the inertial sensors. This frame is denoted by the index b .

In the following, the calculations performed in a strapdown algorithm are addressed. Herein, an n -frame mechanization is assumed. The advantage of an n -frame mechanization is its familiarity, as attitude is given with respect to a local level frame, and velocity is provided in the north, east, and down directions. However, the differential equations in an n -frame mechanization become singular at the Earth's poles; therefore a n -frame mechanization cannot be used if the rover is expected to travel in these regions. When polar travel cannot be precluded, the easiest solution is to use a mechanization in the e -frame, although other alternatives such as the wander azimuth mechanization can be considered as well [28.27].

28.4.2 Attitude Calculations

The attitude of the rover is typically expressed either using Euler angles, a direction cosine matrix (DCM), a quaternion, or the Bortz orientation vector. Attitude representations and their relative tradeoffs have been extensively studied [28.13, 18, 28, 29]. In the attitude calculation implemented in a strapdown algorithm, the differential equations for the selected attitude representation are solved using the angular rate measurements provided by the IMU.

The Euler angles describe the attitude of the body frame with respect to the navigation frame by a series of three rotations. The first rotation is performed through the yaw angle ψ around the z -axis of the navigation frame, which is also referred to as the down axis. The second rotation is performed through the pitch angle θ around the new y -axis. The final rotation through the

roll-angle ϕ around the new x -axis finally leads to the body frame. The Euler angle differential equations are given by

$$\dot{\phi} = (\omega_{nb,y}^b \sin \phi + \omega_{nb,z}^b \cos \phi) \tan \theta + \omega_{nb,x}^b, \quad (28.21)$$

$$\dot{\theta} = \omega_{nb,y}^b \cos \phi - \omega_{nb,z}^b \sin \phi, \quad (28.22)$$

$$\dot{\psi} = \frac{(\omega_{nb,y}^b \sin \phi + \omega_{nb,z}^b \cos \phi)}{\cos \theta}, \quad (28.23)$$

where $\omega_{nb}^b = [\omega_{nb,x}^b, \omega_{nb,y}^b, \omega_{nb,z}^b]^\top$. This set of equations reveals an important drawback. The differential equations for roll and yaw become singular for pitch angles of $\pm 90^\circ$. The reason is that in this case, the first rotation around the z -axis takes place around the same axis as the last rotation around the new x -axis. In other words, the Euler angles become ambiguous in the sense that an arbitrary number of roll-yaw angle pairs can be found that describe a given attitude involving $\pm 90^\circ$ pitch. Therefore, while an attitude involving $\pm 90^\circ$ pitch can be expressed without problems using Euler angles, the differential equation singularity does not allow for an implementation of the attitude calculation in a strapdown algorithm based on Euler angles, for attitudes near $\pm 90^\circ$ pitch. A second disadvantage is that solution of these differential equations require evaluation of several trigonometric functions at the high rate of the IMU gyro integration. The main advantage of Euler angles is their human understandability. However, because the Euler angles can be extracted from the other representations at the low rates at which they are required, the three equations above are rarely implemented.

The DCM or rotation matrix is a 3×3 matrix that defines the transformation of a vector from one coordinate frame to the other. For example, a velocity vector in b -frame coordinates is transformed to n -frame coordinates using the DCM C_b^n as follows

$$v_{eb}^n = C_b^n v_{eb}^b. \quad (28.24)$$

The DCM is an orthonormal matrix, therefore,

$$C_b^n = (C_n^b)^{-1} = (C_n^b)^\top. \quad (28.25)$$

The DCM differential equation is given by

$$\dot{C}_b^n = C_b^n [\omega_{nb}^b \times], \quad (28.26)$$

where

$$[\omega_{nb}^b \times] = \begin{pmatrix} 0 & -\omega_{nb,z}^b & \omega_{nb,y}^b \\ \omega_{nb,z}^b & 0 & -\omega_{nb,x}^b \\ -\omega_{nb,y}^b & \omega_{nb,x}^b & 0 \end{pmatrix} \quad (28.27)$$

is the skew symmetrix matrix corresponding to ω_{nb}^b .

Euler's rotation theorem [28.30] states that any two reference frames related via an arbitrary sequence of rotations can be equivalently represented by a finite rotation about a single axis. This single fixed axis is known as the Euler axis. The Euler axis has the same representation in both reference frames. This demonstrates that the Euler axis is an eigenvector of the DCM corresponding to the eigenvalue 1.

The *Bortz* orientation vector [28.28] is a time-varying vector $\vartheta(t)$ defined by the differential equation

$$\begin{aligned} \dot{\vartheta} &= \omega_{nb}^b + \frac{1}{2} \vartheta \times \omega_{nb}^b \\ &+ \frac{1}{\vartheta^2} \left(1 - \frac{\vartheta \sin(\vartheta)}{2(1 - \cos \vartheta)} \right) \vartheta \times (\vartheta \times \omega_{nb}^b). \end{aligned} \quad (28.28)$$

The length of the orientation vector $\vartheta = \|\vartheta\|$ defines the angle of rotation about the unit vector $\frac{\vartheta}{\vartheta}$ (i. e., Euler axis) such that the navigation frame coincides with the body frame. Therefore, an orientation vector of length ϑ describes the same attitude as the orientation vectors defining the same axis of rotation, but with length $\vartheta \pm 2\pi m$, where m is an arbitrary integer.

The Bortz orientation vector is often the foundation of high-rate attitude integration algorithms designed to address coning [28.10, 11, 31]. For a short period of integration, such that ϑ remains small, (28.28) simplifies to

$$\dot{\vartheta} \approx \omega_{nb}^b + \frac{1}{2} \vartheta \times \omega_{nb}^b + \frac{1}{12} \vartheta \times (\vartheta \times \omega_{nb}^b). \quad (28.29)$$

At the end of the k -th integration period the value of the Bortz orientation vector $\vartheta(\tau_k)$ is saved and the next period of integration starts with zero initial conditions. The sequence of outputs $\vartheta(\tau_k)$ defines a sequence of rotations that can be combined through quaternion multiplication to define the full rotation.

A quaternion [28.29, 32] can be represented as a four element vector, which can be constructed from the Bortz orientation vector as follows

$$q_b^n = \begin{pmatrix} \cos\left(\frac{\vartheta}{2}\right) \\ \left(\frac{\vartheta_x}{\vartheta}\right) \sin\left(\frac{\vartheta}{2}\right) \\ \left(\frac{\vartheta_y}{\vartheta}\right) \sin\left(\frac{\vartheta}{2}\right) \\ \left(\frac{\vartheta_z}{\vartheta}\right) \sin\left(\frac{\vartheta}{2}\right) \end{pmatrix}. \quad (28.30)$$

Note that this is a unit quaternion. Quaternion multiplication can be expressed as a matrix–vector multiplication. For two quaternions, $a = (a_0, a_1, a_2, a_3)$ and

$b = (b_0, b_1, b_2, b_3)$, their product is defined by

$$a \bullet b = \begin{pmatrix} a_0 & -a_1 & -a_2 & -a_3 \\ a_1 & a_0 & -a_3 & a_2 \\ a_2 & a_3 & a_0 & -a_1 \\ a_3 & -a_2 & a_1 & a_0 \end{pmatrix} \bullet \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix}. \quad (28.31)$$

If q_k represents the quaternion corresponding to body rotation $\vartheta(\tau_k)$, then the quaternion for the full rotational sequence is

$$q_0^k = q_k \bullet q_{k-1} \bullet \cdots \bullet q_1. \quad (28.32)$$

Rotations can also be concatenated as follows

$$q_b^n = q_c^n \bullet q_b^c. \quad (28.33)$$

Instead of using the Bortz orientation vector for implementation, the quaternion can be integrated directly. The quaternion differential equation is given by

$$\dot{q}_b^n = \frac{1}{2} q_b^n \bullet \begin{pmatrix} 0 \\ \omega_{nb}^b \end{pmatrix}. \quad (28.34)$$

Strapdown algorithm implementation requires the differential equations for one of the attitude representations to be solved numerically at high rates. Note that the DCM, quaternion, and simplified Bortz ordinary differential equations can be integrated without evaluation of trigonometric functions.

An IMU provides the angular rate vector ω_{ib}^b – or the integral thereof, the angle increments. The ω_{nb}^b is calculated from ω_{ib}^b using

$$\omega_{nb}^b = \omega_{ib}^b - (C_b^n)^\top (\omega_{ie}^n + \omega_{en}^n), \quad (28.35)$$

where the Earth angular rate vector in the n-frame is

$$\omega_{ie}^n = (\Omega \cos \varphi, 0, -\Omega \sin \varphi)^\top. \quad (28.36)$$

Ω is the scalar Earth rate of rotation, and φ denotes the geographic latitude. The term

$$\omega_{en}^n = \begin{pmatrix} +\frac{v_{eb,e}^n}{R_e - h} \\ -\frac{v_{eb,n}^n}{R_n - h} \\ -\frac{v_{eb,e}^n \tan \varphi}{R_e - h} \end{pmatrix} \quad (28.37)$$

is the transport rate, and

$$v_{eb}^n = [v_{eb,n}^n, v_{eb,e}^n, v_{eb,d}^n]^\top.$$

The transport rate takes into account that the navigation frame has to rotate when moving with respect to

Earth's surface, otherwise the z -axis of the n -frame would not maintain its down direction. The Earth's radii of curvature in the north and east directions, R_n and R_e , respectively, are specified in the WGS84 model of the reference ellipsoid.

From the quaternion, the DCM can be calculated using

$$\mathbf{C}_b^n = (\mathbf{C}_b^n(:, 1) \quad \mathbf{C}_b^n(:, 2) \quad \mathbf{C}_b^n(:, 3)), \quad (28.38)$$

with

$$\mathbf{C}_b^n(:, 1) = \begin{pmatrix} (q_0^2 + q_1^2 - q_2^2 - q_3^2) \\ 2(q_1q_2 + q_0q_3) \\ 2(q_1q_3 - q_0q_2) \end{pmatrix}, \quad (28.39)$$

$$\mathbf{C}_b^n(:, 2) = \begin{pmatrix} 2(q_1q_2 - q_0q_3) \\ (q_0^2 - q_1^2 + q_2^2 - q_3^2) \\ 2(q_2q_3 + q_0q_1) \end{pmatrix}, \quad (28.40)$$

$$\mathbf{C}_b^n(:, 3) = \begin{pmatrix} 2(q_1q_3 + q_0q_2) \\ 2(q_2q_3 - q_0q_1) \\ (q_0^2 - q_1^2 - q_2^2 + q_3^2) \end{pmatrix}. \quad (28.41)$$

A rotation from body to the navigation frame is mandatory for velocity calculations in a strapdown algorithm because the specific force is measured in b -frame coordinates whereas velocity and position are required in n -frame coordinates. This rotation can either be accomplished using the DCM or directly from the quaternion. The presentation of the next section assumes that the DCM is computed so that standard matrix–vector notation can be used.

28.4.3 Velocity Calculations

The velocity differential equation in n -frame mechanization is given by

$$\dot{\mathbf{v}}_{eb}^n = \mathbf{C}_b^n \mathbf{f}_{ib}^b - (2\boldsymbol{\omega}_{ie}^n + \boldsymbol{\omega}_{en}^n) \times \mathbf{v}_{eb}^n + \mathbf{g}^n. \quad (28.42)$$

The specific force \mathbf{f}_{ib}^b is provided by the IMU. The multiplication with \mathbf{C}_b^n yields the specific force in n -frame coordinates. Adding the local gravity vector \mathbf{g}^n yields the rover acceleration, (28.12),

$$\mathbf{C}_b^n \mathbf{f}_{ib}^b + \mathbf{g}^n = \mathbf{a}_{ib}^n. \quad (28.43)$$

The local gravity vector is a function of position and can be computed, for example, using the Somigliana formula [28.33]. The term

$$-(2\boldsymbol{\omega}_{ie}^n + \boldsymbol{\omega}_{en}^n) \times \mathbf{v}_{eb}^n$$

is the Coriolis acceleration, an apparent acceleration experienced in rotating coordinate systems, such as the n -frame in which the equation is mechanized. In rotating coordinate systems, a moving object behaves as if this Coriolis acceleration acts on the object. This acceleration cannot be measured by accelerometers; therefore, the Coriolis acceleration has to be added in (28.42). As an example, for the Coriolis acceleration, assume that an object is at rest on the Earth's surface in the equatorial plane. Viewed from an inertial frame, the object trajectory is a circle with a radius that equals the semi-major axis of the Earth ellipsoid. Assume now that an acceleration acts on the object that causes a velocity perpendicular to Earth's surface, so that the height above ground of the object increases. Then, viewed from Earth's surface it is observed that the object is deflected in the west direction by the Coriolis force. In fact, the speed tangential to Earth's surface is not increased; therefore, at a greater distance from the Earth's axis of rotation, the tangential speed of the object is not sufficient to keep up with the point on the Earth's surface from which the object started.

28.4.4 Position Calculations

In the position calculations, the following differential equations are solved numerically

$$\dot{\varphi} = \frac{v_{eb,n}^n}{R_n - h}, \quad (28.44)$$

$$\dot{\lambda} = \frac{v_{eb,e}^n}{(R_e - h) \cos \varphi}, \quad (28.45)$$

$$\dot{h} = v_{eb,d}^n, \quad (28.46)$$

where

$$\mathbf{v}_{eb}^n = [v_{eb,n}^n, v_{eb,e}^n, v_{eb,d}^n]^T.$$

In these equations, the height above ground is negative, because the z -axis of the navigation frame is pointing in the down direction.

Equations (28.42) and (28.44)–(28.46) can be solved numerically, for example, using a Runge–Kutta algorithm [28.34].

28.5 Analysis of Error Effects

The growth of the errors of the INS solution is influenced by the accuracy of the initialization of the strapdown algorithm, the integration algorithm, and by the quality of the inertial sensors. It is useful to distinguish short-term and long-term error characteristics, as is done in the next two sections.

28.5.1 Short-Term Effects

The short-term characteristics are dominant during the first phase of unaided inertial navigation, from initialization or the last aiding measurement time until some minutes afterwards. The impact of the sensor noise was already described in Sect. 28.3.3; the impact of misalignment and SF errors largely depends on the trajectory characteristics. Therefore, the following discussion focuses on the impact of attitude errors and accelerometer and gyroscope biases.

Attitude Errors

Attitude errors lead to errors in the navigation solution, because the specific force is not transformed from b-frame to n-frame coordinates correctly by the first term in (28.42). For example, an acceleration in the north direction measured in b-frame coordinates is not computed as an acceleration in the north direction unless the yaw angle is known perfectly. The effect of attitude errors is more observable when the vehicle is maneuvering; however, even without maneuvering attitude errors can affect the accuracy of the INS position solution. For example, because accelerometers measure the specific force, which is the difference between rover acceleration and gravity, calculation of the rover acceleration requires compensation of gravity. Attitude errors in roll and pitch may lead to an imperfect compensation of the gravity, the uncompensated part is interpreted as rover acceleration and integrated in the SDA, leading to the increase in velocity and position errors. For small angle errors, the uncompensated gravity leads to an erroneous acceleration given by

$$\delta \mathbf{a}_{nb}^b = \begin{pmatrix} -\delta\theta \\ \delta\phi \\ 0 \end{pmatrix} g, \quad (28.47)$$

where $g = |\mathbf{g}^n|$.

Figure 28.5 illustrates a pitch attitude error. In this example, the true pitch angle is zero, whereas the SDA provides a positive, nonzero pitch angle $\delta\theta$. Because of this computed pitch angle, when the SDA computes the navigation frame acceleration by (28.43), the gravity compensation in the x -direction is incorrect by $-g\delta\theta$;

erroneous compensation of the gravity leads to an erroneous acceleration in the negative body x -direction (28.47). The down direction is less affected by this mechanism, as the erroneous gravity compensation in down direction is proportional to

$$1 - \cos(\delta\phi) \cos(\delta\theta),$$

which is close to zero for small angle errors.

A similar assessment can be made for a roll angle error. Therefore, in the short-term, roll and pitch attitude errors cause horizontal acceleration error, leading to a linear growth of velocity errors and a quadratic growth of position errors with time, whereas the vertical channel is less sensitive to attitude errors.

Gyroscope Biases

As long as the attitude errors are sufficiently small, gyroscope biases lead to a linear growth of attitude errors with time. Combined with the results from the previous paragraph, it can be concluded that this causes an acceleration error in the horizontal channels that grows linearly with time, leading to a quadratic growth of the velocity errors and cubic growth of the position errors.

Accelerometer Biases

An uncompensated accelerometer bias is interpreted as a rover acceleration, leading to a linear growth of velocity errors and quadratic growth of position errors with time.

Summary

From the above discussion, it is clear that INS errors grow at rates faster than linear growth. Note that it is possible for horizontal accelerometer bias errors to exactly cancel attitude errors. When this occurs, the errors have no effect, and are said to lie in the unobservable space. Aiding the INS with other sensors tends to drive these errors to the unobservable space.

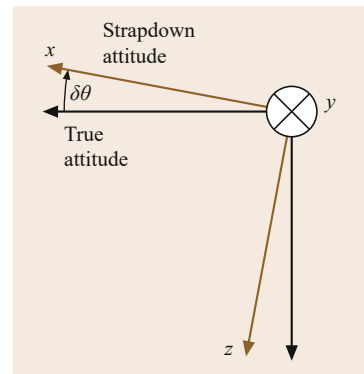


Fig. 28.5 Illustration of a pitch error

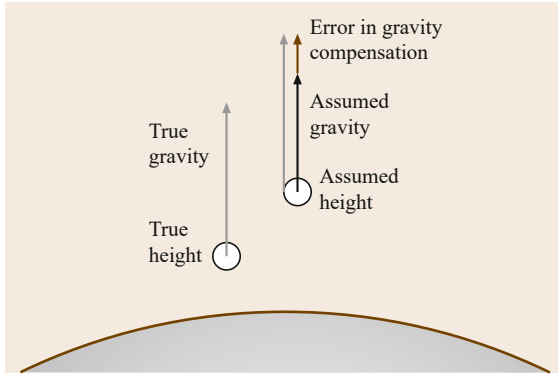


Fig. 28.6 INS vertical channel instability

28.5.2 Long-Term Effects

For the long-term error characteristics of an INS, the instability of the vertical channel and the Schuler oscillations become important.

Instability of the Vertical Channel

Computation of acceleration from the measured specific force vector requires gravity compensation. This compensation computes the expected gravity at the estimated position using a suitable model (e.g., Somigliana gravity formula). The magnitude of the gravity vector reduces with increasing height above ground. Because the strapdown algorithm provides an erroneous height, the expected magnitude of the gravity vector does not match the true gravity vector. For example, when the computed height of the rover above ground is higher than the true position (Fig. 28.6), the computed magnitude of the gravity vector is too small. Therefore, only a part of the gravity contributing to the measured specific force is compensated, whereas the remaining part is interpreted erroneously as an acceleration of the rover. Unfortunately, the vertical channel error dynamics are unstable; therefore, this erroneous acceleration increases the present height error.

Due to this instability, an initial height error of only 10 m causes a height error of several thousand meters after 1 h of unaided inertial navigation. Therefore, any INS that operates autonomously for longer periods of time requires aiding of the vertical channel, for example, using a barometric altimeter, for stabilization.

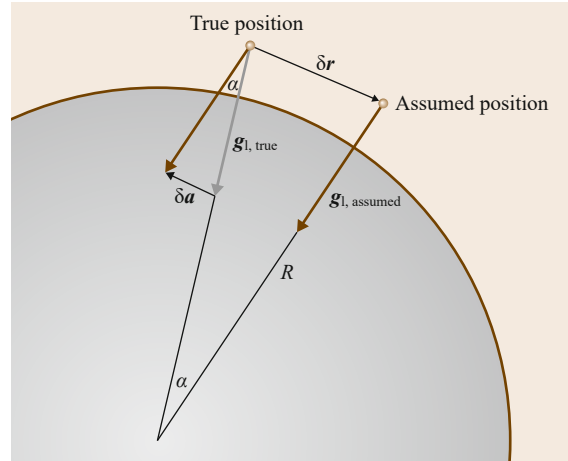


Fig. 28.7 Illustration of the Schuler oscillation mechanism

Schuler Oscillations

Although the gravitational errors destabilize the vertical channel, they provide a stabilizing effect on the horizontal channels, resulting in *Schuler oscillations* [28.35].

The mechanism that leads to Schuler oscillations is illustrated in Fig. 28.7. Due to a horizontal position error $\delta \mathbf{r}$, the strapdown algorithm computes a slightly erroneous direction of the gravity vector. Therefore, the gravity measured with the specific force is not compensated correctly in the velocity calculation of the strapdown algorithm. The uncompensated portion is interpreted erroneously as rover acceleration. Fortunately, this acceleration error $\delta \mathbf{a}$ is pointing in the direction toward the true rover position, that is, acting against the position error. From Fig. 28.7, the following approximate relation is obtained

$$\frac{\delta \mathbf{r}}{R} = -\frac{\delta \mathbf{a}}{g} . \quad (28.48)$$

This leads to

$$\delta \ddot{\mathbf{r}} + \frac{g}{R} \delta \mathbf{r} = 0 , \quad (28.49)$$

which is the differential equation of a harmonic oscillator. The resulting Schuler oscillations of the position and velocity errors of the INS solution show a period of approximately 84 min.

28.6 Aided Navigation

The INS achieves a high-bandwidth rover state vector estimate $\hat{\mathbf{x}}(t)$ by propagating (28.7) through time using the function f described in Sect. 28.4. The time propagation only uses the IMU sensor data.

At time t_j , the INS provides an estimate $\hat{\mathbf{z}}_j = \hat{\mathbf{z}}(t_j)$ of the vector $\mathbf{z}_j = \mathbf{z}(t_j)$. Section 28.2.1 discussed the definition of the vector \mathbf{z} that depends on the rover state \mathbf{x} and the sensor calibration parameters \mathbf{c}_u and \mathbf{c}_y . The navigation system also maintains an estimate of the error covariance matrix \mathbf{P}_{z_j} .

The definition, causes, and effects of INS state errors are discussed in Sects. 28.2.4, 28.3.3, and 28.5. The rate of error accumulation is dependent on the

quality of the IMU and the accuracy of $\hat{\mathbf{c}}_u$. The need for aiding of the INS by additional sensors is clear. Each aiding sensor measurement provides information useful for correcting the rover state vector $\hat{\mathbf{x}}$ and the IMU calibration parameters $\hat{\mathbf{c}}_u$, decreasing the INS error vector $\delta\mathbf{z}$ in certain directions of the state space. In the process, the accuracy of the aiding sensor calibration vector $\hat{\mathbf{c}}_y$ will also be enhanced.

This process is called state estimation. The subsequent sections will consider INS aiding (i.e., state estimation) in various forms using GNSS observables as the aiding signals.

28.7 State Estimation

This section briefly introduces the main state estimation algorithm. This algorithm can be derived from multiple perspectives, maximum a posteriori (MAP) or minimum mean squared error [28.14, 16, 36–38] (Chap. 22).

For a linear system with appropriate assumptions, the optimal algorithm is known as the *Kalman filter* [28.39]. Because the INS time propagation and measurement models are nonlinear, the extended Kalman filter is often utilized in aided inertial navigation applications. A few alternative estimation approaches are discussed later in Sect. 28.10.

The extended Kalman filter (EKF) state estimator uses the new information in the measurement to compute an improved, ideally optimal, posterior state estimate $\hat{\mathbf{z}}_j^+$ with error covariance $\mathbf{P}_{z_j}^+$. In this notation, an upper index $\{.\}^+$ denotes a quantity in which the measurement information has already been incorporated, known as an a posteriori quantity, and an upper index $\{.\}^-$ denotes an a priori quantity, for which the measurement information has not yet been incorporated.

Assume that at t_j , an aiding measurement $\tilde{\mathbf{y}}_j = \tilde{\mathbf{y}}(t_j)$ is available. Using the INS state estimate $\hat{\mathbf{z}}_j^-$, the INS predicts the value of the aiding measurement $\hat{\mathbf{y}}_j^-$ using (28.8). The residual measurement is

$$\mathbf{r}_j = \tilde{\mathbf{y}}_j - \hat{\mathbf{y}}_j^- . \quad (28.50)$$

Under the assumptions that linearization errors are small enough to be ignored and that

$$\delta\mathbf{z}(t_j) \sim N(\mathbf{0}, \mathbf{P}_{z_j}^-),$$

where $\mathbf{P}_{z_j}^-$ is the error covariance matrix computed by the INS, then

$$\mathbf{r}_j \sim N(\mathbf{0}, \mathbf{S}_j) ,$$

where

$$\mathbf{S}_j = \mathbf{H}_j \mathbf{P}_{z_j}^- \mathbf{H}_j^\top + \mathbf{R}_j , \quad (28.51)$$

$$\mathbf{H}_j = \left. \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \right|_{\mathbf{z}=\hat{\mathbf{z}}_j^-} . \quad (28.52)$$

The fact that $\mathbf{r}_j \sim N(\mathbf{0}, \mathbf{S}_j)$ with both \mathbf{r}_j and \mathbf{S}_j computed on-line allows the measurement validity to be evaluated in real-time using the chi-squared variable $\mathbf{r}_j^\top \mathbf{S}_j^{-1} \mathbf{r}_j$ (Chap. 24).

When the measurement is deemed to be valid, the improved state estimate and its error covariance are computed as

$$\hat{\mathbf{z}}_j^+ = \hat{\mathbf{z}}_j^- + \mathbf{K}_j \mathbf{r}_j , \quad (28.53)$$

$$\mathbf{P}_{z_j}^+ = \mathbf{P}_{z_j}^- - \mathbf{K}_j \mathbf{S}_j \mathbf{K}_j^\top , \quad (28.54)$$

where

$$\mathbf{K}_j = \mathbf{P}_{z_j}^- \mathbf{H}_j^\top (\mathbf{H}_j \mathbf{P}_{z_j}^- \mathbf{H}_j^\top + \mathbf{R}_j)^{-1} .$$

If the measurements are deemed to be invalid, then the corrections due to those measurements would not be applied (equivalently, $\mathbf{K}_j = \mathbf{0}$).

Equation (28.54) shows that aiding measurements decrease the error covariance in a direction determined by \mathbf{K}_j . Any single measurement will only correct a certain linear subspace of the state space, which leaves

the error to accumulate in the complementary subspace, which is said to be unobservable from this measurement. As long as each subspace of the state space is corrected sufficiently often, the overall state estimate remains accurate. The error covariance matrix keeps track of these details over time.

It is well known [28.20] that for an INS observability is dependent on the rover motion. For example, certain subspaces that contain the attitude errors and

IMU biases will be unobservable from position or velocity aiding, during time intervals when the rover is nonaccelerating. This means that the estimation error will accumulate in those unobservable subspaces. It also means that the accumulated errors will have no effect on the measured outputs during the period they are not observable. Depending on the amount of error accumulation, their effect can be significant when the rover motion causes the errors to become observable.

28.8 GNSS and Aided INS

This section presents an overview of a few typical GNSS-aided INS formulations. Each subsection presents one approach along with a brief discussion. The discussion and figures should be interpreted in a general sense. Designers have implemented many variations on the general ideas described herein.

To allow a straightforward comparison of approaches, this section will neglect lever arm corrections. The lever arm is the vector from the GNSS antenna to the IMU effective location. When the lever arm is significant relative to the accuracy that the system is designed to achieve, lever arm compensation must be taken into account. A detailed discussion of lever arm corrections is presented in Sect. 28.9.2.

28.8.1 Loose (Position Domain) Coupling

Figure 28.8 shows a typical block diagram for a loosely coupled approach. The GNSS receiver uses the observed satellite signals to compute the *measurement* of the GNSS antenna position $\tilde{\mathbf{p}}$ and velocity $\tilde{\mathbf{v}}$. The INS uses its state vector $\hat{\mathbf{z}}$ to compute the estimate of the GNSS antenna position $\hat{\mathbf{p}}$ and velocity $\hat{\mathbf{v}}$. The EKF uses the position and velocity residual vectors along with the error models to estimate the error state $\delta\hat{\mathbf{z}}$, which is fed

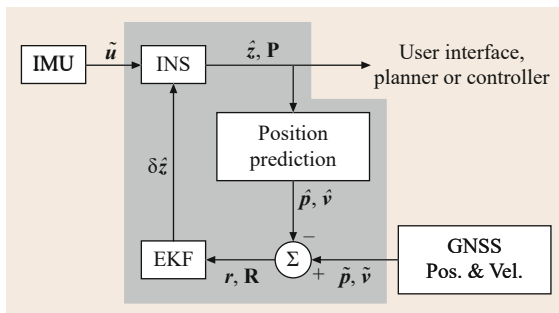


Fig. 28.8 GNSS loosely coupled INS aiding block diagram

back into the INS, which removes the estimated errors from the system.

Loosely coupled approaches require the least amount of GNSS knowledge by the navigation system designer. The designer need not understand the ephemeris computations, clock models, receiver specific implementation issues, or various other items discussed in the GNSS interface documentation.

Various factors come into play related to the computation of the GNSS position and velocity *measurements* by the receiver. These factors are manufacturer dependent and may not be fully understandable or controllable by the user:

- Position and velocity are not the basic measurements of a GNSS receiver. Instead, position and velocity can be computed by the receiver at any time when signals from at least four satellites are available. The accuracy of the position and velocity solutions depends on the number and geometry of the available satellites. This accuracy can change significantly over time. Some receivers output quantities such as horizontal dilution of precision (HDOP), GDOP, and so on, which are useful, but they do not allow complete determination of the measurement covariance matrix \mathbf{R} (28.4) that is used by the EKF to determine the optimal gain \mathbf{K}_i . When the complete \mathbf{R} matrix is unknown, the design is forced to fill in the missing knowledge with assumptions that are not necessarily correct, resulting in a suboptimal design.
- The receiver may include internal filters (Sect. 28.10.1). Such filters are problematic for INS aiding, as the EKF in the INS is designed based on the assumption that the error (or noise) $\boldsymbol{\eta}_y$ in (28.4) is white; however, the GNSS receiver filters may make this assumption invalid.

Without an internal navigation filter, the GNSS receiver will not produce any position and velocity outputs

at times when signals from fewer than four satellite vehicles are available. At these time instants, the information from the available satellites is lost. The internal filter is often preferred by GPS-only users of the receiver, but is a complication that can cause anomalous behavior in GNSS-aided INS implementations. For INS aiding, the receiver settings should eliminate its internal navigation filter.

28.8.2 Tight (Observable Domain) Coupling

Figure 28.9 shows a typical block diagram for a tightly coupled approach. The GNSS receiver tracks the radio frequency signals using phase-lock and delay-lock loops to extract information to determine the measurements of pseudorange $\hat{\rho}$, Doppler \hat{D} , and carrier-phase $\hat{\phi}$, along with various signal quality indicators and satellite ephemeris data (Chaps. 14 and 15). The INS uses its estimated state \hat{z} along with the satellite ephemeris data and equations describing the satellite orbits (Chap. 3) to predict the pseudorange $\hat{\rho}$, Doppler \hat{D} , and carrier phase $\hat{\phi}$. The EKF uses some combination of the pseudorange, Doppler, and carrier-phase residuals to estimate the error state $\delta\hat{z}$, which is fed back into the INS to remove the estimated errors from the system.

The tightly coupled approach requires more effort and knowledge on the part of the designer and more computation by the INS computer. The INS must implement the satellite ephemeris equations to compute the satellite position and velocity necessary for predicting the pseudorange, Doppler, and carrier-phase. The designer must also understand certain intricacies (e.g., clock models and interpretations of quality indicators) of the GNSS receiver with which they are working. The return on the invested design effort is the opportunity for improved performance. INS aiding is possible by whichever satellites are available, even when the number available is less than four. The accuracy of each satellite measurement can be independently and accurately characterized. In addition, validity decisions can be made on a per satellite basis, before they effect the navigation solution.

If the GNSS receiver contains an internal navigation filter, that filter should have no effect on the pseudorange, Doppler, and carrier-phase measurements. The receiver does contain filters in the radio frequency stage to reduce noise and radio interference. The radio frequency filters have center frequencies and bandwidths that are sufficiently high that they do not adversely affect INS aiding. The receiver also contains filters within the receiver code tracking delay-lock-loop (DLL) and carrier tracking phase-lock-loops (PLLs). The trade-offs – involving accuracy, loss-of-lock, and other issues – in selecting these bandwidths are intricate and

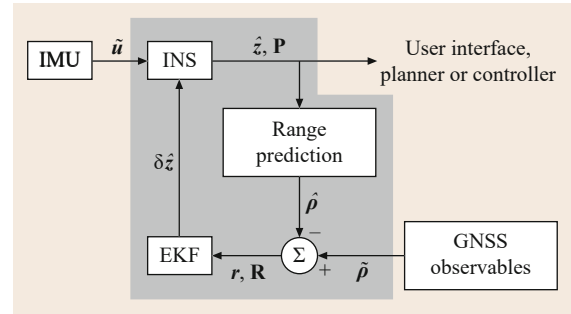


Fig. 28.9 GNSS tightly coupled INS aiding block diagram

are determined by the receiver designers. The measurement errors introduced by the DLL and PLL are independent for each satellite; however, depending on the receiver, the time correlation may be significant enough that it needs to be taken into account in the INS error model.

At present, most available GNSS receivers, if provided with differential correction signals, will utilize the corrections in its computed position, but will not account for the corrections in the pseudorange, Doppler, and carrier-phase measurements that are output. The designer must account for the differential corrections in the residual formation process.

28.8.3 Ultra-Tight or Deep Coupling

Figure 28.10 shows an example block diagram for an ultratightly or deeply coupled approach. The INS uses its estimated state \hat{z} along with the satellite ephemeris data and equations describing the satellite orbits to predict the pseudorange $\hat{\rho}$, Doppler \hat{D} , and carrier-phase $\hat{\phi}$. These predictions are injected into the GNSS receiver carrier-phase and/or code tracking algorithms. The receiver carrier-phase and code tracking errors then serve directly as residuals for the EKF to estimate the error state $\delta\hat{z}$, which is fed back into the INS to remove the estimated errors from the system [28.40, 41].

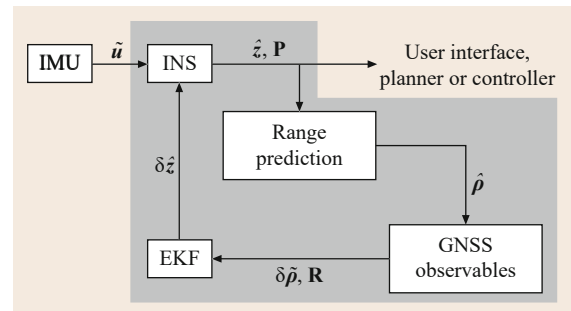


Fig. 28.10 GNSS ultra or deeply coupled INS aiding INS block diagram



Fig. 28.11 GPS position fixes obtained during a test drive (courtesy of Stadt Karlsruhe, Liegenschaftsamt)

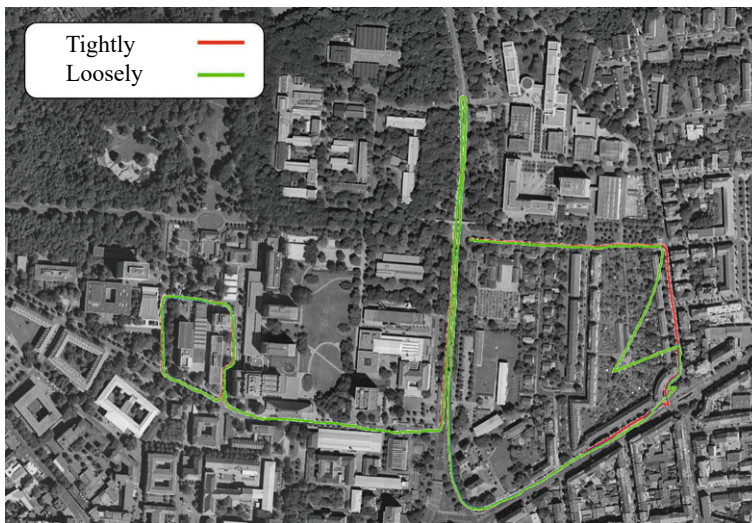


Fig. 28.12 Navigation solutions of a loosely and tightly coupled GPS/INS system obtained by post-processing of recorded GPS and IMU data (courtesy of Stadt Karlsruhe, Liegenschaftsamt)

In contrast to the independent receiver tracking loops of the tightly and loosely coupled approaches, for the approaches depicted in Fig. 28.10, the code and carrier replica generation process within the receiver for all satellites is driven by the INS solution.

In each of Figs. 28.8, 28.9, and 28.10, the gray background indicates the portion of the system software that will be the concern of the INS designers. The gray background for Figs. 28.8 and 28.9 looks similar; however, the GPS processing portion of the software in the tightly coupled approach of Fig. 28.9 will be more involved than for the loosely coupled approach. The system software for the ultratightly coupled approach is significantly more involved and includes alteration of the GNSS internal receiver software (or firmware) that implements the carrier-phase and code tracking loops.

Design of an ultratightly or deeply coupled system typically requires collaboration between the INS and GNSS receiver design teams.

The advantages of the ultratightly coupled approach include the following. Accurate prediction of the satellite range and Doppler by the INS allows for faster and weaker signal acquisition. The INS allows the motion of the receiver antenna that is within the IMU bandwidth to be accounted for by the INS, outside the GPS receiver signal-tracking software; therefore, the bandwidth of the receiver tracking loops can be significantly decreased, which has benefits for noise reduction and lowering susceptibility to receiver jamming. Because the full vehicle bandwidth is within the bandwidth of the IMU, the approach also yields better satellite signal tracking during highly dynamic maneuvers. These factors generate en-

hanced accuracy, availability, and continuity at the expense of more advanced onboard processing.

28.8.4 Illustrative Comparison

To illustrate the performance differences between loosely and tightly coupled integration architectures, navigation solutions are compared that were obtained by a post-processing of GPS and MEMS IMU data that was recorded during a test drive. Post-processing assures that in both integration architectures, the identical sensor data was processed.

Figure 28.11 shows the position fixes provided by the GPS receiver during the test drive. On the right, a significant gap is visible between the position fixes. This was due to the fact that while driving along this street, the surrounding buildings blocked the view to some of the GPS satellites, so that the number of available satellites dropped below four.

Figure 28.12 shows the position solution using loose (green) and tight (red) integration architectures.

Most of the time, architectures provide results that are indistinguishable at the scale of the image. However, during the aforementioned section of the test drive when less than four satellites are visible, the strap-down solution of the loosely coupled system remains unaided, as no GPS position fixes was available. Consequently, the position solution of the loosely coupled system drifts away from the true position (moving along the road). The error accumulation is smooth and its rate is determined by the quality of the IMU and the INS implementation. Once the GPS receiver is able to compute position again, the accumulated error is corrected and the state estimate snaps back to near the correct location. For the tightly coupled system, the pseudorange measurements from the two to three satellites that were still visible in this section of the test drive provided enough information to maintain an accurate position solution. Note that as the vehicle maneuvers, the actual satellites available at each epoch may change to correct the state in different directions.

28.9 Detailed Example

In the following, the design of a navigation filter for tight integration of GNSS and INS is sketched.

28.9.1 System Model

For the example design, the state vector

$$\mathbf{z} = (\mathbf{p}^n, \mathbf{v}_{\text{eb}}^n, \mathbf{q}_b^n, \mathbf{b}_a, \mathbf{b}_\omega, \mathbf{b}_c)^\top \quad (28.55)$$

contains the rover position, velocity, attitude quaternion, accelerometer biases, gyroscope biases, and the receiver clock error vector \mathbf{b}_c . The vector \mathbf{b}_c is composed of the receiver clock ct_r in meters and clock rate $\dot{c}t_r$ in m/s. This is a typical set of choices, which yields good performance for almost all applications and IMU grades. However, in some cases, it might be advantageous to add additional states to account for IMU SF and misalignment errors, time correlated errors such as those induced by vibrations, or time correlated GNSS ranging errors (e.g., code multipath).

The navigation system usually has two major components: the INS and the error estimator. The INS performs temporal updates by integrating the IMU data through the kinematic model to maintain the INS *total state vector* $\hat{\mathbf{z}}$. As the INS integrates, for the various reasons already discussed, the uncertainty in the error between \mathbf{z} and $\hat{\mathbf{z}}$ increases. The error estimator (Sects. 28.6 and 28.7) uses the aiding information to

perform measurement updates that compute an estimate $\delta\hat{\mathbf{z}}$ of the *error state vector*.

For this design example, the aiding information derives from GPS measurements and the error state vector is defined as

$$\delta\mathbf{z} = (\delta\mathbf{p}^n, \delta\mathbf{v}_{\text{eb}}^n, \boldsymbol{\psi}_n^{\hat{n}}, \delta\mathbf{b}_a, \delta\mathbf{b}_\omega, \delta\mathbf{b}_c)^\top.$$

In this definition, $\delta\mathbf{p}^n$ are the position errors in the directions north, east, and down; $\delta\mathbf{v}_{\text{eb}}^n$ are the velocity errors in n -frame coordinates; $\boldsymbol{\psi}_n^{\hat{n}}$ is the three dimensional attitude error vector; $\delta\mathbf{b}_a$ and $\delta\mathbf{b}_\omega$ are the errors of the accelerometer and gyroscope biases, respectively; and $\delta\mathbf{b}_c$ is the vector composed of the receiver clock error $c\delta t_r$ in meters and clock error drift rate $c\delta\dot{t}_r$ in m/s.

The estimated error state $\delta\hat{\mathbf{z}}$ is then used to correct the total system state $\hat{\mathbf{z}}$. After this correction, the estimated error state vector is reset to zero. Such a configuration is known as error state feedback formulation, or error state closed loop formulation.

For the time propagation, the total system state is propagated by the SDA using IMU data. The error state time propagation is trivial. Having been set to zero after correcting the total states, the estimated error state remains zero when propagated forward in time. Therefore, the time propagation of the error state vector is not implemented. The error estimation algorithm does

propagate the error covariance matrix through time, by (28.11).

Note that the choice of reference frames for the various components of the total and error state vectors is completely independent. Similarly, the choice of attitude representations for the total and error state vectors is an independent choice. In this example, the attitude error vector is not a quaternion. The interpretation of the attitude error vector will be further discussed following (28.59).

The filter design requires a differential equation describing the dynamics of the error state. This differential equation is defined by differencing (28.1) and (28.6) and transforming to the appropriate reference frame. These equations are derived subsequently. To keep the derivations concise, terms which are not essential have been neglected. For example, for a system which processes GNSS measurements typically every second, but at least every few minutes, Coriolis error terms are small; velocity errors can be assumed to not be influenced by the position errors; and, the unit vector pointing from the GNSS antenna to a satellite can also be assumed to not be affected by position errors. All these simplifications have negligible impact on the filter performance. The neglected terms would become significant in the description of the long-term behavior of the navigation errors. Frequent updates from the GNSS measurements render these terms unnecessary.

Position Error Differential Equation

The position error differential equations are derived from (28.44)–(28.46) for latitude, longitude and height, respectively. From these equations, it is obvious that the time derivative of the latitude depends on north velocity, height, and, via the radius of curvature in the north direction, also on latitude. Similar dependencies occur in longitude equation. However, following the reasoning of the previous section, the dependencies on position are weak and can be safely neglected. The resulting position error differential equation is

$$\delta \dot{\mathbf{p}}^n = \delta \mathbf{v}_{\text{eb}}^n. \quad (28.56)$$

Velocity Error Differential Equation

In the derivation of the velocity error differential equations, the Coriolis term in the velocity differential equation is neglected, leading to

$$\dot{\mathbf{v}}_{\text{eb}}^n \approx \mathbf{C}_{\text{b}}^n \dot{\mathbf{f}}_{\text{ib}}^{\text{b}} + \mathbf{g}^n. \quad (28.57)$$

The equation for the estimated quantities can be similarly formulated as

$$\hat{\dot{\mathbf{v}}}_{\text{eb}}^n \approx \mathbf{C}_{\text{b}}^{\hat{n}} \hat{\dot{\mathbf{f}}}_{\text{ib}}^{\text{b}} + \hat{\mathbf{g}}^n. \quad (28.58)$$

We choose to model the relationship between the true and estimated attitude as

$$\mathbf{C}_{\text{b}}^{\hat{n}} = \mathbf{C}_{\text{n}}^{\hat{n}} \mathbf{C}_{\text{b}}^{\text{n}}. \quad (28.59)$$

Hereby, the attitude errors have been allocated to the navigation frame, the difference between true and estimated navigation frame is represented by the DCM $\mathbf{C}_{\text{n}}^{\hat{n}}$. The attitude errors are assumed to be small; therefore, this DCM is expressed as

$$\mathbf{C}_{\text{n}}^{\hat{n}} = (\mathbf{I} + [\boldsymbol{\psi}_{\text{n}}^{\hat{n}} \times]) = (\mathbf{I} + \boldsymbol{\Psi}_{\text{n}}^{\hat{n}}), \quad (28.60)$$

where

$$\boldsymbol{\psi}_{\text{n}}^{\hat{n}} = (\alpha, \beta, \gamma)^{\top} \quad (28.61)$$

is the vector of attitude errors. This representation of the DCM is obtained by expressing the DCM as a sequence of small angle rotations and using the small angle approximations

$$\cos(\Delta) \approx 1 \text{ and } \sin(\Delta) \approx \Delta$$

in the DCM definition. Therefore, α , β , and γ can be interpreted as small angle errors around the navigation frame north, east, and down axes.

Substituting (28.60) into (28.59) leads to

$$\mathbf{C}_{\text{b}}^{\hat{n}} = (\mathbf{I} + \boldsymbol{\Psi}_{\text{n}}^{\hat{n}}) \mathbf{C}_{\text{b}}^{\text{n}}, \quad (28.62)$$

and

$$\mathbf{C}_{\text{b}}^{\text{n}} = (\mathbf{I} - \boldsymbol{\Psi}_{\text{n}}^{\hat{n}}) \mathbf{C}_{\text{b}}^{\hat{n}}, \quad (28.63)$$

which follows from the orthonormality of the DCM and the fact that transposing a skew symmetric matrix changes the signs of the off-diagonal elements.

Defining the relationship between true and estimated specific force as

$$\hat{\mathbf{f}}_{\text{ib}}^{\text{b}} = \mathbf{f}_{\text{ib}}^{\text{b}} + \delta \mathbf{f}_{\text{ib}}^{\text{b}}, \quad (28.64)$$

the time derivative of the velocity errors is obtained by subtracting (28.57) from (28.58)

$$\begin{aligned} \delta \dot{\mathbf{v}}_{\text{eb}}^n &= \dot{\mathbf{v}}_{\text{eb}}^n - \dot{\hat{\mathbf{v}}}_{\text{eb}}^n \\ &= \mathbf{C}_{\text{b}}^{\hat{n}} \hat{\dot{\mathbf{f}}}_{\text{ib}}^{\text{b}} - (\mathbf{I} - \boldsymbol{\Psi}_{\text{n}}^{\hat{n}}) \mathbf{C}_{\text{b}}^{\hat{n}} (\hat{\mathbf{f}}_{\text{ib}}^{\text{b}} - \delta \mathbf{f}_{\text{ib}}^{\text{b}}) \\ &= \boldsymbol{\Psi}_{\text{n}}^{\hat{n}} \mathbf{C}_{\text{b}}^{\hat{n}} \hat{\dot{\mathbf{f}}}_{\text{ib}}^{\text{b}} + \mathbf{C}_{\text{b}}^{\hat{n}} \delta \dot{\mathbf{f}}_{\text{ib}}^{\text{b}} \\ &= -\left[\mathbf{C}_{\text{b}}^{\hat{n}} \hat{\mathbf{f}}_{\text{ib}}^{\text{b}} \times \right] \boldsymbol{\Psi}_{\text{n}}^{\hat{n}} + \mathbf{C}_{\text{b}}^{\hat{n}} \delta \dot{\mathbf{f}}_{\text{ib}}^{\text{b}}. \end{aligned} \quad (28.65)$$

Now, the relationship between errors in specific force and quantities contained in the error state vector has to be established. The measured specific force is given by

$$\hat{f}_{ib}^b = f_{ib}^b + b_a + n_a, \quad (28.66)$$

where f_{ib}^b is the true specific force, b_a are the accelerometer biases, and n_a is the noise corrupting the accelerometer measurements. If additional sensor errors like SFs or misalignments shall be estimated, this equation has to be augmented accordingly. An estimate of the specific force can be calculated from the measured specific force as follows

$$\hat{f}_{ib}^b = \tilde{f}_{ib}^b - \hat{b}_a. \quad (28.67)$$

This leads to

$$\begin{aligned} \delta f_{ib}^b &= \hat{f}_{ib}^b - f_{ib}^b \\ &= (\tilde{f}_{ib}^b - \hat{b}_a) - f_{ib}^b \\ &= (f_{ib}^b + b_a + n_a - \hat{b}_a) - f_{ib}^b \\ \delta f_{ib}^b &= -\delta b_a + n_a, \end{aligned} \quad (28.68)$$

and finally

$$\delta \dot{v}_{eb}^n = [\hat{f}_{ib}^n \times] \psi_n^{\hat{n}} - C_b^{\hat{n}} \delta b_a + C_b^{\hat{n}} n_a, \quad (28.69)$$

where

$$\hat{f}_{ib}^{\hat{n}} = C_b^{\hat{n}} \hat{f}_{ib}^b.$$

In this equation, the first two terms on the right hand side are time-varying linear functions of the error state vector. The last term shows how the (random) accelerometer measurement noise drives the error state. This will be discussed further in the summary portion of this section.

Attitude Error Differential Equation

A simplified attitude error differential equation is obtained from an approximation of the DCM differential equation given by

$$\dot{C}_b^n = C_b^n \Omega_{nb}^b \approx C_b^n \Omega_{ib}^b. \quad (28.70)$$

Again, a similar equation can be formulated for the estimated quantities

$$\dot{C}_b^{\hat{n}} = C_b^{\hat{n}} \hat{\Omega}_{nb}^b \approx C_b^{\hat{n}} \hat{\Omega}_{ib}^b. \quad (28.71)$$

The time derivative of (28.59) is given by

$$\begin{aligned} \dot{C}_b^{\hat{n}} &= \frac{d}{dt} (C_b^{\hat{n}} C_b^n) \\ &= \dot{C}_b^{\hat{n}} C_b^n + C_b^{\hat{n}} \dot{C}_b^n \\ &= \dot{\Psi}_n^{\hat{n}} C_b^n + C_b^{\hat{n}} \dot{C}_b^n. \end{aligned} \quad (28.72)$$

Inserting (28.71) yields

$$\begin{aligned} C_b^{\hat{n}} \hat{\Omega}_{ib}^b &\approx \dot{\Psi}_n^{\hat{n}} C_b^n + C_b^{\hat{n}} C_b^n \Omega_{ib}^b, \\ C_b^{\hat{n}} \hat{\Omega}_{ib}^b &\approx \dot{\Psi}_n^{\hat{n}} C_b^n + C_b^{\hat{n}} \Omega_{ib}^b, \end{aligned} \quad (28.73)$$

and finally (dropping second-order error terms)

$$\begin{aligned} \dot{\Psi}_n^{\hat{n}} C_b^n &\approx C_b^{\hat{n}} (\hat{\Omega}_{ib}^b - \Omega_{ib}^b), \\ \dot{\Psi}_n^{\hat{n}} &\approx C_b^{\hat{n}} \delta \Omega_{ib}^b (C_b^n)^\top, \\ \dot{\Psi}_n^{\hat{n}} &\approx C_b^{\hat{n}} \delta \Omega_{ib}^b (C_b^{\hat{n}})^\top (I + \Psi_n^{\hat{n}}), \\ \dot{\Psi}_n^{\hat{n}} &\approx C_b^{\hat{n}} \delta \Omega_{ib}^b (C_b^{\hat{n}})^\top. \end{aligned} \quad (28.74)$$

This can be expressed without skew symmetric matrices

$$\dot{\Psi}_n^{\hat{n}} \approx C_b^{\hat{n}} \delta \omega_{ib}^b. \quad (28.75)$$

A derivation similar to that for (28.67) yields

$$\delta \omega_{ib}^b = -\delta b_\omega + n_\omega, \quad (28.76)$$

which leads to the desired attitude error differential equation

$$\dot{\Psi}_n^{\hat{n}} \approx -C_b^{\hat{n}} \delta b_\omega + C_b^{\hat{n}} n_\omega. \quad (28.77)$$

Similar to the velocity error differential equation, the first term is a time-varying linear functions of the error state vector. The last term shows how the (random) gyro measurement noise drives the error state. Such driving noise terms are referred to as *process noise*.

Clock Error Differential Equation

A simple model for the receiver clock error is

$$\delta \dot{t}_c = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \delta b_c + \begin{pmatrix} v_{c\delta t} \\ v_{c\delta i} \end{pmatrix}, \quad (28.78)$$

where $v_{cT} = [v_{c\delta t}, v_{c\delta i}]^\top$ is a white Gaussian noise vector process [28.42].

Summary of Error Model

Summing up, the navigation filter system model in continuous time is given by

$$\delta \dot{\mathbf{z}} = \mathbf{F} \delta \mathbf{z} + \mathbf{G} \mathbf{v}. \quad (28.79)$$

The system matrix is given by

$$\mathbf{F} = \begin{pmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{F}_{23} & -\mathbf{C}_b^{\hat{n}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{C}_b^{\hat{n}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{F}_{66} \end{pmatrix}, \quad (28.80)$$

where

$$\mathbf{F}_{23} = -[\hat{\mathbf{f}}_{ib}^{\hat{n}} \times] \text{ and } \mathbf{F}_{66} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \quad (28.81)$$

The process noise mapping matrix \mathbf{G} is

$$\mathbf{G} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{C}_b^{\hat{n}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{C}_b^{\hat{n}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix}, \quad (28.82)$$

where the process noise vector is

$$\mathbf{v} = (\mathbf{n}_a, \mathbf{n}_\omega, \mathbf{v}_{b_a}, \mathbf{v}_{b_\omega}, \mathbf{v}_{cT})^\top. \quad (28.83)$$

In these equations, the inertial sensor biases are modeled as random walk processes with driving noise \mathbf{v}_{b_a} and \mathbf{v}_{b_ω} . Alternatively, the biases can be modeled by various other Gauss–Markov processes, the modification of the system matrix for this is straightforward.

Often, the error estimation filter is implemented via the Kalman filter. In addition to the above definitions of \mathbf{F} and \mathbf{G} , the Kalman filter requires specification of the PSD matrix \mathbf{Q} for the process noise vector \mathbf{w} . By (28.83) and the fact that the noise sources are independent, the matrix \mathbf{Q} is block diagonal with five blocks. The four IMU blocks are determined from the Allan variance specifications for the IMU provided by its manufacturer. Specification of the block for the clock process noise is determined by the Allan variance parameters of the receiver clock. This is distinct from many other Kalman filter applications in which Kalman filter *tuning* is a time-consuming art.

28.9.2 Measurement Models

To complete this tightly coupled design example, models for GPS observables are required. We only include discussion of pseudorange and Doppler observables. Carrier-phase aiding can be achieved by similar methods, but is not discussed herein due to the complexities related to integer ambiguity resolution [28.43–46]. It should be noted that the measurements reported as Doppler measurements by many receivers are actually delta ranges computed over short time intervals. In such cases, more accurate estimation may be possible, especially during periods of acceleration, by more accurate models of the delta range observable; however, the resulting algorithms are more complicated to implement.

Because the IMU and GNSS antenna cannot be physically co-located, the position, velocity, and acceleration of these two items are distinct. Depending on the expected rover dynamics and the distance between the IMU and GPS antenna, it is sometimes necessary to consider this lever arm, \mathbf{l}^b , in the aiding equations.

Influence of the Leverarm

The position of the GNSS antenna, \mathbf{p}_A^n , is related to the rover position, \mathbf{p}_U^n (defined by the intersection of the sensitive axes of the inertial sensors), as follows

$$\mathbf{p}_A^n = \mathbf{p}_U^n + \mathbf{C}_b^n \mathbf{l}^b. \quad (28.84)$$

With (28.63) this leads to

$$\mathbf{p}_A^n = \mathbf{p}_U^n + (\mathbf{I} - \Psi_n^{\hat{n}}) \mathbf{C}_b^{\hat{n}} \mathbf{l}^b. \quad (28.85)$$

The estimated antenna position is

$$\hat{\mathbf{p}}_A^n = \hat{\mathbf{p}}_U^n + \mathbf{C}_b^{\hat{n}} \mathbf{l}^b. \quad (28.86)$$

The error in rover position and rover attitude relate to the error in antenna position

$$\begin{aligned} \delta \mathbf{p}_A^n &= \hat{\mathbf{p}}_A^n - \mathbf{p}_A^n \\ &= \delta \hat{\mathbf{p}}_U^n + \Psi_n^{\hat{n}} \mathbf{C}_b^{\hat{n}} \mathbf{l}^b \\ &= \delta \hat{\mathbf{p}}_U^n - [\mathbf{C}_b^{\hat{n}} \mathbf{l}^b \times] \Psi_n^{\hat{n}}. \end{aligned} \quad (28.87)$$

Equation (28.87) would also be the basis for processing position measurements in a loosely coupled system.

Similarly, the GNSS antenna velocity differs from the rover velocity depending on the length of the lever arm and the angular rate. With (28.63) and using the analog of the (28.67) for angular rates, the relationship between the velocity of the GNSS antenna and the rover can be expressed as

$$\begin{aligned} \mathbf{v}_{eA}^n &= \mathbf{v}_{eb}^n + \mathbf{C}_b^n (\boldsymbol{\omega}_{eb}^b \times \mathbf{l}^b) \\ &= \mathbf{v}_{eb}^n + (\mathbf{I} - \Psi_n^{\hat{n}}) \mathbf{C}_b^{\hat{n}} \\ &\quad ((\boldsymbol{\omega}_{eb}^b - \delta \boldsymbol{\omega}_{eb}^b) \times \mathbf{l}^b). \end{aligned} \quad (28.88)$$

An estimate of the GNSS antenna velocity is

$$\hat{\mathbf{v}}_{\text{eA}}^{\text{n}} = \hat{\mathbf{v}}_{\text{eb}}^{\text{n}} + \mathbf{C}_{\text{b}}^{\hat{\text{n}}} (\hat{\boldsymbol{\omega}}_{\text{eb}}^{\text{b}} \times \mathbf{l}^{\text{b}}), \quad (28.89)$$

leading to the following relationship between the errors in GNSS antenna velocity, rover velocity, rover attitude, and gyroscope biases

$$\begin{aligned} \delta \mathbf{v}_{\text{eA}}^{\text{n}} &= \hat{\mathbf{v}}_{\text{eA}}^{\text{n}} - \mathbf{v}_{\text{eA}}^{\text{n}} \\ &= \delta \mathbf{v}_{\text{eb}}^{\text{n}} + \boldsymbol{\Psi}_{\text{n}}^{\hat{\text{n}}} \mathbf{C}_{\text{b}}^{\hat{\text{n}}} \hat{\boldsymbol{\Omega}}_{\text{eb}}^{\text{b}} \mathbf{l}^{\text{b}} + \mathbf{C}_{\text{b}}^{\hat{\text{n}}} \delta \boldsymbol{\Omega}_{\text{eb}}^{\text{b}} \mathbf{l}^{\text{b}} \\ &= \delta \mathbf{v}_{\text{eb}}^{\text{n}} - \left[\mathbf{C}_{\text{b}}^{\hat{\text{n}}} \hat{\boldsymbol{\Omega}}_{\text{eb}}^{\text{b}} \mathbf{l}^{\text{b}} \times \right] \boldsymbol{\Psi}_{\text{n}}^{\hat{\text{n}}} - \mathbf{C}_{\text{b}}^{\hat{\text{n}}} [\mathbf{l}^{\text{b}} \times] \delta \boldsymbol{\omega}_{\text{eb}}^{\text{b}} \\ &\approx \delta \mathbf{v}_{\text{eb}}^{\text{n}} - \left[\hat{\boldsymbol{\Omega}}_{\text{ib}}^{\text{n}} \mathbf{l}^{\text{n}} \times \right] \boldsymbol{\Psi}_{\text{n}}^{\hat{\text{n}}} + \mathbf{C}_{\text{b}}^{\hat{\text{n}}} [\mathbf{l}^{\text{b}} \times] \delta \mathbf{b}_{\omega}, \end{aligned} \quad (28.90)$$

where

$$\mathbf{l}^{\text{n}} = \mathbf{C}_{\text{b}}^{\hat{\text{n}}} \mathbf{l}^{\text{b}}.$$

Equation (28.90) is also the basis for processing velocity measurements in a loosely coupled system.

With these provisions, the pseudorange and delta-range measurement models can be derived.

Pseudorange Measurement

As described in Chap. 19, neglecting common mode errors (e.g., ephemeris, ionosphere, troposphere, and satellite clock), the measurement of the pseudorange ρ_s to satellite s can be modeled as

$$\tilde{\rho}_s = \|\mathbf{p}_s^{\text{n}} - \mathbf{p}_{\text{A}}^{\text{n}}\| + c\delta t_r + v_{\rho}, \quad (28.91)$$

where \mathbf{p}_s^{n} denotes the satellite position in navigation frame coordinates and v_{ρ} is the noise corrupting the pseudorange measurement. The total differential of (28.91) is

$$\delta \tilde{\rho}_s = \frac{\partial \rho_s}{\partial \mathbf{p}_{\text{A}}^{\text{n}}} \delta \mathbf{p}_{\text{A}}^{\text{n}} + \frac{\partial \rho_s}{\partial \mathbf{b}_c} \delta \mathbf{b}_c, \quad (28.92)$$

where $\delta \mathbf{b}_c$ was defined in (28.55) and the derivatives evaluate to

$$\frac{\partial \tilde{\rho}_s}{\partial \mathbf{b}_c} = [1 \ 0] \quad \text{and} \quad \frac{\partial \tilde{\rho}_s}{\partial \mathbf{p}_{\text{A}}^{\text{n}}} = -[\mathbf{e}_s^{\text{n}}]^{\top}, \quad (28.93)$$

where \mathbf{e}_s^{n} is the unit line-of-sight vector between the antenna and satellite.

The measurement matrix \mathbf{H}_{ρ} can be formulated using submatrices which describe the relationship between a pseudorange measurement and components of the filter state vector as

$$\mathbf{H}_{\rho} = (\mathbf{H}_{\rho,p}, \mathbf{0}, \mathbf{H}_{\rho,\psi}, \mathbf{0}, \mathbf{0}, \mathbf{H}_{\rho,ct_r}). \quad (28.94)$$

From (28.87) and (28.93), the submatrices of the pseudorange measurement matrix are

$$\mathbf{H}_{\rho,p} = -[\mathbf{e}_s^{\text{n}}]^{\top}, \quad (28.95)$$

$$\mathbf{H}_{\rho,\psi} = [\mathbf{C}_{\text{b}}^{\hat{\text{n}}} \mathbf{l}^{\text{b}} \times], \quad (28.96)$$

$$\mathbf{H}_{\rho,ct_r} = [1 \ 0]. \quad (28.97)$$

All other submatrices are zero vectors.

The residual measurement δy processed in an error state space formulation is the difference between the actual measurement and the measurement prediction. For the pseudorange measurement, this evaluates to

$$\delta y_{\rho} = \tilde{\rho}_i - \hat{\rho}_i. \quad (28.98)$$

Doppler Measurement

A Doppler measurement provides the relative velocity between GNSS antenna and satellite, projected onto the line-of-sight toward the satellite. As this is a measurement of the Doppler offset of the satellite signal, the frequency error of the receiver clock enters the delta-range measurement model

$$\dot{\rho}_s = \mathbf{e}_s^{\text{n}} \cdot (\mathbf{v}_{\text{es}}^{\text{n}} - \mathbf{v}_{\text{eA}}^{\text{n}}) + c\delta \dot{t}_r. \quad (28.99)$$

The total differential of (28.99) is

$$\delta \dot{\rho}_s = \frac{\partial \dot{\rho}_s}{\partial \mathbf{v}_{\text{eA}}^{\text{n}}} \delta \mathbf{v}_{\text{eA}}^{\text{n}} + \frac{\partial \dot{\rho}_s}{\partial \mathbf{b}_c} \delta \mathbf{b}_c, \quad (28.100)$$

where the derivatives evaluate to

$$\frac{\partial \dot{\rho}_s}{\partial \mathbf{v}_{\text{eA}}^{\text{n}}} = -(\mathbf{e}_s^{\text{n}})^{\top}, \quad (28.101)$$

and

$$\frac{\partial \dot{\rho}_s}{\partial \mathbf{b}_c} = (0 \ 1). \quad (28.102)$$

Again, the Kalman filter measurement matrix can be formulated using submatrices as follows

$$\mathbf{H}_{\dot{\rho}} = (\mathbf{0}, \mathbf{H}_{\dot{\rho},v}, \mathbf{H}_{\dot{\rho},\psi}, \mathbf{0}, \mathbf{H}_{\dot{\rho},b_{\omega}}, \mathbf{H}_{\dot{\rho},ct_r}). \quad (28.103)$$

From the derivatives in (28.101), (28.102), and (28.90), the submatrices are

$$\mathbf{H}_{\dot{\rho},v} = -(\mathbf{e}_s^{\text{n}})^{\top}, \quad (28.104)$$

$$\mathbf{H}_{\dot{\rho},\psi} = (\mathbf{e}_s^{\text{n}})^{\top} [\hat{\boldsymbol{\Omega}}_{\text{ib}}^{\text{n}} \mathbf{l}^{\text{n}} \times], \quad (28.105)$$

$$\mathbf{H}_{\dot{\rho},b_{\omega}} = -(\mathbf{e}_s^{\text{n}})^{\top} \mathbf{C}_{\text{b}}^{\hat{\text{n}}} [\mathbf{l}^{\text{b}} \times], \quad (28.106)$$

$$\mathbf{H}_{\dot{\rho},ct_r} = (0 \ 1). \quad (28.107)$$

All other submatrices are zero. The residual measurement δy processed in an error state space formulation for the Doppler measurement is

$$\delta y_{\dot{\rho}} = \tilde{\rho}_i - \mathbf{e}_s^n \cdot (\mathbf{v}_{es}^n - \hat{\mathbf{v}}_{eA}^n) - c\delta\hat{t}_r, \quad (28.108)$$

where $\hat{\mathbf{v}}_{eA}^n$ is given by (28.89) and $\tilde{\rho}_i$ represents the Doppler measurement.

Navigation Correction

After processing the pseudorange and deltarange measurements available at the current epoch, the Kalman filter state vector contains estimated errors, which are used in closed-loop operation to correct the total quantities, thereby implementing *tight integration*.

Latitude, longitude, and height are corrected, respectively, using the following equations

$$\hat{\phi}^+ = \hat{\phi}^- - \frac{\delta\hat{x}_n^+}{R_n - \hat{h}^-}, \quad (28.109)$$

$$\hat{\lambda}^+ = \hat{\lambda}^- - \frac{\delta\hat{x}_e^+}{(R_e - \hat{h}^-) \cos \hat{\phi}^-}, \quad (28.110)$$

$$\hat{h}^+ = \hat{h}^- - \delta\hat{x}_d^+. \quad (28.111)$$

The velocity is corrected using

$$\hat{\mathbf{v}}_{eb}^{n,+} = \hat{\mathbf{v}}_{eb}^{n,-} - \delta\hat{\mathbf{v}}_{eb}^{n,+}. \quad (28.112)$$

From the estimated attitude error vector, a correction quaternion

$$\boldsymbol{\sigma}_c = -\delta\hat{\boldsymbol{\psi}}, \quad (28.113)$$

$$\boldsymbol{\sigma}_c = \sqrt{\boldsymbol{\sigma}_c^\top \boldsymbol{\sigma}_c}, \quad (28.114)$$

$$\mathbf{q}_c = \begin{pmatrix} \cos \frac{\sigma_c}{2} \\ \frac{\sigma_c}{2} \sin \frac{\sigma_c}{2} \end{pmatrix} \quad (28.115)$$

is calculated, which is used to correct the strapdown attitude quaternion

$$\hat{\mathbf{q}}_b^{n,+} = \mathbf{q}_c \cdot \hat{\mathbf{q}}_b^{n,-}. \quad (28.116)$$

Finally, the biases and the receiver clock error estimates are corrected

$$\hat{\mathbf{b}}_a^+ = \hat{\mathbf{b}}_a^- - \delta\hat{\mathbf{b}}_a^+, \quad (28.117)$$

$$\hat{\mathbf{b}}_\omega^+ = \hat{\mathbf{b}}_\omega^- - \delta\hat{\mathbf{b}}_\omega^+, \quad (28.118)$$

$$\hat{\mathbf{b}}_c^+ = \hat{\mathbf{b}}_c^- - c\delta\hat{t}_r. \quad (28.119)$$

As mentioned previously, after the estimated errors have been used to correct the total quantities, the error filter state vector is set to zero

$$\delta\mathbf{z}^+ = \mathbf{0} \quad (28.120)$$

to avoid these estimated errors are being used again at the next measurement step.

Implementation Aspects

When implementing an aided INS, some subtleties are important.

Sequential Measurement Processing. At each epoch, a varying number of pseudorange and deltarange measurements are available. For the Kalman filter processing of these measurements, two options exist. One option is to form a vector of all measurements, which is then processed in one single measurement update. The other option is to perform a scalar measurement update for each pseudorange and each deltarange measurement. The motivation for such sequential scalar measurement processing is that in the calculation of the Kalman gain matrix, using the standard EKF measurement update in (28.53) and (28.54), the matrix inversion reduces to a simple division. If all measurements are processed in one single step, a matrix with the number of rows equal to the number of available pseudoranges and deltaranges has to be inverted. This inversion is possible, but computationally more demanding. Alternatively, computation of the Kalman gain matrix can be avoided altogether by instead solving an appropriately weighted set of linear equations containing the prior and the measurement residuals [28.47, 48].

Time of Measurement Validity. The pseudorange and deltarange measurements are available for processing only after some delay (i. e., latency), typically between 50 and 200 ms. This delay is due to the time required for the receiver internal signal processing and the time required to transfer the data from the receiver to the navigation processor. Especially for high-speed or high-dynamics applications, this latency has to be considered. A very simple technique which usually delivers a close-to-optimal performance is to store the total navigation state at the time of validity of the measurements, and then to calculate the measurement residual for the Kalman filter when the measurements become available not based on the current rover state, but based on the rover state at the time of measurement validity. Denoting the current IMU epoch with the index k and assuming that the measurements were valid n IMU epochs in the past, the Kalman filter state vector update in the error state space formulation is given with this technique by

$$\delta\mathbf{z}_k^+ = \delta\mathbf{z}_k^- + \mathbf{K}_k [(\hat{\mathbf{y}}_{k-n} - \tilde{\mathbf{y}}_{k-n}) - \mathbf{H}_k \delta\mathbf{z}_k^-]. \quad (28.121)$$

Note that $\delta\mathbf{z}_k$ in this equation is really the correction valid at $(k-n)$. The time propagation of the error state vector from $(k-n)$ to k has been neglected. For higher fidelity, this time propagation is easily accomplished using n iterations of the error state transition matrix as defined in (28.10).

Detection of Outliers. It should be noted that prior to processing any type of measurement, a health check should be performed to identify and reject outliers (aiding measurements, IMU data, and model mismatch).

A simple but effective approach to detect outliers in the aiding measurements is to calculate the squared weighted norm of the residual vector \mathbf{r} and compare it to a threshold. The observed measurement vector is then considered unlikely, and ignored, when

$$\|\mathbf{r}\|_{\mathbf{S}}^2 = \mathbf{r}^T \mathbf{S}^{-1} \mathbf{r} \geq \chi_{\alpha}^2(q, 0), \quad (28.122)$$

where

$$\mathbf{r} = \tilde{\mathbf{y}} - \hat{\mathbf{y}},$$

with covariance

$$\mathbf{S} = (\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}),$$

and $\chi_{\alpha}^2(q, 0)$ is the threshold computed from the central Chi-squared distribution with $q = \dim(\mathbf{r})$ degrees of freedom (Chap. 24).

Outlier detection using the squared Mahalanobis distance $\|\mathbf{r}\|_{\mathbf{S}}^2$ only works well when the covariance matrices \mathbf{P} and \mathbf{R} are accurate statistical representations for the actual errors of the state estimate and measurement errors. The Kalman filter designer should always check, in simulations, that performance is adequate under diverse conditions. A conservative filter tuning (i. e., \mathbf{P} is larger than the covariances of the actual errors), is – within some range – usually not such a big problem, maybe some accuracy is lost; however, \mathbf{P} being too small can cause the measurements to be weighted too little; the filter relies too much on its state estimate, and finally diverges. To check that a filter is implemented properly, especially, the quantities that are not measured directly (e.g., attitude and inertial sensor biases) need to be investigated. The position and velocity errors in a GNSS/INS system are largely dominated by the accuracy of the GNSS measurements, so flaws in the implemented system model might not be visible from these errors, whereas they clearly show in the attitude and bias estimates.

Yaw Angle Observability. A system is said to be observable, if it is possible to estimate the system state from the available measurements [28.19]. For example, consider a simple system where the system state contains position and velocity. If a series of position measurements is available, the system is observable, because position is measured directly and velocity can be inferred from the time history of position measurements. If only velocity measurements are available, the system is unobservable: velocity is measured directly, but it is not possible to obtain an estimate of position from velocity measurements only.

For GNSS/INS systems, observability depends on the trajectory characteristics. For a moment, to allow a straightforward discussion, assume that the accelerometer biases are zero. We also drop the additive noise, as it has no effect on observability. According to (28.69), the relationship between a change in velocity errors and the attitude errors is

$$\frac{d}{dt} \begin{pmatrix} \delta v_n \\ \delta v_e \\ \delta v_d \end{pmatrix} = \begin{pmatrix} 0 & \hat{f}_d & -\hat{f}_e \\ -\hat{f}_d & 0 & \hat{f}_n \\ \hat{f}_e & -\hat{f}_n & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}, \quad (28.123)$$

where \hat{f}_n , \hat{f}_e , and \hat{f}_d are the components of the specific force vector in the north, east, and down directions, respectively. Looking at the specific force skew symmetric matrix, it is found that a yaw error γ can only affect the velocity error vector when the specific force in east or north directions is nonzero, respectively. The velocity errors are observable from position and/or velocity measurements, or from a sufficiently diverse set of pseudorange and/or deltarange measurements in the case of a tightly coupled system. However, when the horizontal specific force components (f_n, f_e) are zero, the yaw angle error cannot affect the velocity errors, and is therefore unobservable. As a consequence, for any GNSS/INS system that is at rest or in nonaccelerating motion (e.g., straight and level flight), the uncertainty in the yaw angle error will grow with time. If such a situation is expected, two options exist, either the accuracy of the gyros must be sufficient so that the growth in yaw error with time is not a problem for the expected duration without horizontal accelerations, or additional sensors have to be used which provide yaw angle observability, for example, a magnetometer or feature sensor.

Equation (28.123) would seem to imply that the tilt errors (α, β) are observable due to \hat{f}_d being nonzero because of gravity; however, this is not the case. It is only an artifact of the simplifying assumption stated to derive that equation. When the biases are not zero, analysis starting from the original equations shows that the tilt errors cannot be distinguished from the accelerometer bias errors, unless the specific force vector changes sufficiently (e.g., acceleration occurs). However, position or velocity measurements drive the linear combination of tilt and accelerometer bias errors to an unobservable space, on which they have no effect on those position and velocity measurements [28.20, 49, 50].

In free fall, the specific force is zero, so the attitude would be unobservable from position and velocity measurements. Similarly, for a satellite in orbit, the specific force is zero, so for a GNSS/INS system onboard a satellite, attitude would be completely unobservable. For satellites, star trackers are frequently used to achieve attitude observability.

28.10 Alternative Estimation Methods

This section briefly discusses various alternative GNSSs related state estimation approaches in comparison to the GNSS-aided INS approach that has been the main topic of this chapter.

28.10.1 Standalone GNSS

The topic of this section is state estimation based only upon GNSS, without IMU. This topic is thoroughly covered in Chap. 21 so this discussion only includes the detail necessary to allow comparison with GNSS-aided INS. The discussion will focus on pseudorange processing, but the ideas directly extend to carrier-phase and Doppler processing.

Any time at which the GNSS receiver has measured pseudorange $\tilde{\rho} \in \mathbb{R}^{n_s}$ from at least $n_s \geq 4$ satellites, it can solve

$$\tilde{\rho} = \mathbf{h}(\mathbf{x}) + \eta_\rho, \quad (28.124)$$

for $\mathbf{x} = [\mathbf{p}, ct_r] \in \mathbb{R}^4$, where \mathbf{p} is the GNSS receiver antenna position, ct_r is the receiver clock bias, and $\eta_\rho \sim N(\mathbf{0}, \mathbf{R})$.

Equation (28.124) is often solved using iterative weighted least squares (WLS). Starting with an estimate $\hat{\mathbf{x}}$, the receiver predicts

$$\hat{\rho} = \mathbf{h}(\hat{\mathbf{x}})$$

and computes the residual measurement

$$\delta\rho = \tilde{\rho} - \hat{\rho}.$$

The correction $\delta\mathbf{x}$ satisfies the normal equation

$$\mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H} \delta\mathbf{x} = \mathbf{H}^\top \mathbf{R}^{-1} \delta\rho, \quad (28.125)$$

where $\mathbf{H} \in \mathbb{R}^{n_s \times 4}$ is defined in (28.52). Given the solution $\delta\mathbf{x}$, the estimate is corrected as

$$\hat{\mathbf{x}} = \hat{\mathbf{x}} + \delta\mathbf{x}. \quad (28.126)$$

This process of measurement prediction and state correction repeats, using the same measurement, until $\delta\mathbf{x}$ is sufficiently small. This solution method fails when the rank of \mathbf{H} is less than four.

Even though the noise on each of the individual pseudorange measurements is uncorrelated with the noise on other pseudorange measurements at a given time instant, the solution of (28.125) results in the estimated state vector having cross-correlated components. The error covariance of the solution is

$$\text{cov}(\hat{\mathbf{x}}) = (\mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H})^{-1}. \quad (28.127)$$

Most receivers do not output this matrix, especially the off-diagonal elements. Instead, receivers use certain portions of the diagonal of this matrix to compute HDOP, position dilution of precision (PDOP), vertical dilution of precision (VDOP), and so on. Lack of access to this matrix complicates the design of loosely coupled systems.

The pseudorange measurement error η_ρ accounts for various error sources. Depending on the application scenario, the error sources can include atmospheric delay, ephemeris model errors, multipath, and noise. Because the first three of these errors are time correlated, the position estimation error will also be time correlated. Equation (28.127) accounts only for the error at a single time instant, not the time correlation of the error. For a stationary receiver, the correlation time of the multipath error is significant over minutes. In such a scenario, faster sampling does not allow averaging to decrease the effects of such errors. For a moving receiver, the time correlation properties of the multipath effects are highly variable.

To enable the GNSS receiver to output a position estimate even when $n_s < 4$ and to filter noise, many GNSS receivers include a navigation filter. The filter time propagation equations are

$$\hat{\mathbf{x}}_{k+1}^- = \Phi \hat{\mathbf{x}}_k^+, \quad (28.128)$$

$$\hat{\mathbf{P}}_{k+1}^- = \Phi \mathbf{P}_k^+ \Phi^\top + \mathbf{Q}_d. \quad (28.129)$$

The filter measurement updates equations are

$$\hat{\mathbf{K}}_k = \hat{\mathbf{P}}_k^- \hat{\mathbf{H}}_k^\top (\hat{\mathbf{H}}_k \hat{\mathbf{P}}_k^- \hat{\mathbf{H}}_k^\top + \mathbf{R}_k)^{-1}, \quad (28.130)$$

$$\hat{\mathbf{P}}_k^+ = (\mathbf{I} - \hat{\mathbf{K}}_k \hat{\mathbf{H}}_k) \hat{\mathbf{P}}_k^-, \quad (28.131)$$

$$\hat{\mathbf{x}}_k^+ = \hat{\mathbf{x}}_k^- + \hat{\mathbf{K}}_k (\tilde{\rho}_k - \hat{\rho}_k). \quad (28.132)$$

These equations implement a Kalman filter. However, it is difficult to claim that this Kalman filter has any optimality properties relative to an actual system assumed to be modeled by

$$\mathbf{x}_{k+1} = \Phi \mathbf{x}_k + \mathbf{v}_k, \quad (28.133)$$

because there is no physical basis for defining the filter parameters Φ and $\mathbf{Q}_d = \text{cov}(\mathbf{v}_k)$ at the point in time when the receiver is designed. A typical method to address this issue is to allow the user some flexibility in the specification of Φ and \mathbf{Q}_d . Three typical options are the P, PV, and PVA models.

Position

In cases where the position (P) is modeled as a random walk

$$\Phi = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \Phi_c \end{bmatrix}, \quad (28.134)$$

$$\mathbf{x} = [\mathbf{p}^\top, \mathbf{x}_c^\top]^\top, \quad (28.135)$$

$$\mathbf{Q}_d = \text{diag}([\mathbf{Q}_p, \mathbf{Q}_c]), \quad (28.136)$$

where the receiver manufacturer specifies the clock model parameters Φ_c and \mathbf{Q}_c . This model assumes that the antenna velocity can be accurately modeled as a white random process with constant covariance \mathbf{Q}_p for all times.

Position–Velocity (PV)

When the receiver is in uniform motion (i.e., slowly varying velocity), performance is improved by including velocity states in the model

$$\Phi = \begin{bmatrix} \mathbf{I} & T\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Phi_c \end{bmatrix}, \quad (28.137)$$

$$\mathbf{x} = [\mathbf{p}^\top, \mathbf{v}^\top, \mathbf{x}_c^\top]^\top, \quad (28.138)$$

$$\mathbf{Q}_d = \text{diag}([\mathbf{Q}_{pv}, \mathbf{Q}_c]), \quad (28.139)$$

where T is the time step between GNSS measurements. This model assumes that the antenna acceleration can be accurately modeled as a white random process yielding a constant covariance \mathbf{Q}_{pv} for all times. Ideally, \mathbf{Q}_{pv} would be selected to accurately quantify the *random* variation in \mathbf{v} over all times; however, this is not feasible given that in most applications this variation is not a stationary random process. Therefore, in reality, specification of the value of \mathbf{Q}_{pv} is a tuning process for each specific application. Since a stationary stochastic model is being *fit* to a nonstationary process, a rigorous specification approach cannot be expected.

Position–Velocity–Acceleration (PVA)

When the velocity cannot reasonably be modeled to be constant, then an acceleration state can be added in each of the three orthogonal directions. Instead of a pure random walk, the acceleration is usually modeled as a scalar Gauss–Markov process (i.e., low frequency, colored noise). This model is reasonable because accelerations are not usually constant, but are correlated over short time intervals. The resulting dynamic model

has

$$\Phi = \begin{bmatrix} \mathbf{I} & T\mathbf{I} & \mu_1\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mu_2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mu_3\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Phi_c \end{bmatrix} \quad (28.140)$$

$$\mathbf{x} = [\mathbf{p}^\top, \mathbf{v}^\top, \mathbf{a}^\top, \mathbf{x}_c^\top]^\top \quad (28.141)$$

$$\mathbf{Q}_d = \text{diag}([\mathbf{Q}_{pva}, \mathbf{Q}_c]) \quad (28.142)$$

where

$$\lambda > 0,$$

$$\mu_1 = \frac{(\mu_3 - 1 + \lambda T)}{\lambda^2},$$

$$\mu_2 = \frac{(1 - \mu_3)}{\lambda}, \quad \text{and}$$

$$\mu_3 = \exp(-\lambda T).$$

The process driving noise \mathbf{v}_k accounts for the *random* variations in the acceleration. Ideally, \mathbf{Q}_{pva} would be selected to accurately quantify the *random* variation in \mathbf{a} over all times; however, as stated above for \mathbf{Q}_{pv} , this is not feasible due to the nonstationarity of the acceleration process. Instead, a value of \mathbf{Q}_{pva} is selected that provides a reasonable (i.e., compromise) performance under the expected range of acceleration conditions.

Comparison

The purpose of this section is to discuss a few points important in the design of GNSS-aided INS systems.

First, the Kalman filter in the GNSS receiver is designed by tuning the model choice (e.g., P, PV, and PVA) and parameters \mathbf{Q}_d to yield a satisfactory performance tradeoff. In all cases, there is no rigorous method to select \mathbf{Q}_p , \mathbf{Q}_{pv} , or \mathbf{Q}_{pva} . The receiver may allow some user choices, such as high, medium, or low dynamics; however, there should be no misconception that this is an accurate model of the dynamic process over all times. The resulting receiver navigation filter is not optimal and has no stochastic interpretations. The filter gain \mathbf{K} could just as well have been designed by other methods, such as pole-placement. In particular, the matrix \mathbf{P} does not indicate the covariance of the estimation error.

Second, if $n_s < 4$, then the navigation filter can still incorporate the information from the available measurements using (28.130)–(28.132). This corrects the state in certain directions while allowing the uncertainty to increase in the other directions.

Third, over time intervals without GNSS measurements, the state trajectory can be estimated by iterating (28.128) and (28.129). However, this prediction should

be considered carefully. The P-model will predict the current position estimate forward in time, regardless of the actual antenna motion. The PV-model will make a linear prediction using the current position and velocity estimates, regardless of the actual antenna motion. The PVA-model will make a parabolic prediction using the current position, velocity, and acceleration estimates, regardless of the actual antenna motion.

Using the GNSS receiver's filtered output to aid an INS in such a situation is clearly inappropriate, because the IMU provides the INS with information unavailable to the GNSS receiver. Even when $n_s \geq 4$, the GNSS receiver's internal navigation filter is problematic for INS aiding applications: the GNSS navigation internal filter should be disabled. The INS computation of the vehicle trajectory, based on the IMU inputs using (28.7), will be much more responsive than the estimates of the GNSS receiver's internal filter. The bandwidth of the INS trajectory estimation is determined by the bandwidth of the IMU and in a proper design is much higher than the bandwidth of the vehicle motion. The bandwidth of the GNSS receiver trajectory estimate is determined by the bandwidth of the receiver tracking loops, the receiver sample time, and the hypothesized receiver navigation filter parameters Φ and \mathbf{Q}_d . The GNSS receiver bandwidth is typically much lower than the INS or vehicle bandwidth. Taking GNSS measurements at a higher rate does not alleviate the issues.

28.10.2 Advanced Bayesian Estimation

The MAP estimate of the trajectory

$$\mathbf{Z} = \{\mathbf{z}(t) \text{ for } t = t_0, \dots, t_K\}$$

maximizes the cost function [28.37, 51–53]

$$J(\mathbf{Z}) = (p_{z_0}(\mathbf{z}(t_0))p_{\eta_y}(\mathbf{Y} - \mathbf{h}(\mathbf{Z}))p_{v_u}(\mathbf{Z}_+ - \phi(\mathbf{Z}, \mathbf{U})|\mathbf{z}_0)), \quad (28.143)$$

where

$$\mathbf{z}(t_0) \sim N(\mathbf{z}_0, \mathbf{P}_0)$$

is the prior distribution of the initial state,

$$\begin{aligned} \mathbf{Z}_+ &= \{\mathbf{z}(t) \text{ for } t = t_1, \dots, t_K\}, \\ \mathbf{Y} &= \{\mathbf{y}(t) \text{ for } t = t_1, \dots, t_M\} \end{aligned}$$

is the set of all GNSS measurements from the initial to the present time, $\mathbf{U} = \{\mathbf{u}(t) \text{ for } t = \tau_1, \dots, \tau_K\}$ is the set of IMU measurements from the initial to the present time, and the operator ϕ is defined in (28.7). This formulation assumes that $\mathbf{x}(t_0)$, \mathbf{v}_u , \mathbf{v}_{c_u} , \mathbf{v}_{c_y} , and η_y are all mutually independent.

Direct maximization of (28.143) is complicated by various factors. First, each $\mathbf{x}(t_k) \in \mathbb{R}^n$ and $n \geq 15$. Second, the number of time instances K grows without bound as the duration of an application increases. In addition, the IMU time τ is small, the total number of GNSS measurements M increases with K , and the number of GNSS measurements per time step is time-varying. Third, the kinematic model \mathbf{f} and the measurement models \mathbf{h} are nonlinear. Fourth, in general the various probability density functions can be non-Gaussian and multimodal.

The EKF, which has been the focus of this chapter, is only one approximate solution of the nonlinear Bayesian estimation problem. The EKF is designed based on linearized error models and uses the assumption that the various noise sources are normally distributed. Normal noise distributions are reasonably valid for GPS and IMU sensors. Linearized error models are discussed further below.

Particle and Unscented Filtering

The particle filter (PF) and unscented Kalman filter are two other approximate solutions that have become popular. The PF approximates the probability density function of the state estimate with a set of probability-weighted particles. Each particle is an instance of the state vector. Each particle is propagated through time using the kinematic model. The probabilistic weights are adjusted using the measurement model. Detailed descriptions of each of these steps are presented in [28.14, 38]. The PF is expected to be more accurate than the EKF when the EKF linearization and normal density assumptions are not valid. For example, if the density is multimodal or if the model second derivatives are significant relative to the uncertainty of the state estimate. The computational cost of the PF is related to the number of particles. The number of particles required for an accurate representation of the density function grows exponentially with the dimension of the state vector. Due to the INS state vector having dimension 15 or larger, the number of particles is large enough that PFs typically are not used.

Once the state is initialized, the state error covariance is typically small enough that the neglected second derivatives of the EKF approach have little impact; therefore, a main issue is getting past the initialization step. In particular, the accuracy of the attitude is important, especially yaw. Traditional means to initialize the state for military applications include in-flight alignment or gyro compassing. Lower cost and commercial applications typically use GNSS and an electronic compass, with the accelerometer serving as an inclinometer at initialization. Another approach is to use real-time smoothing as outlined in the next subsection.

Realtime Nonlinear Bayesian Estimation

Recent advances in the field robotics literature for the real-time solution of the simultaneous location and mapping (SLAM) problem using a formulation similar to that of (28.143) have interesting utility in the GNSS-aided INS field.

The likelihood function corresponding to the cost function of (28.143) reduces to the nonlinear least squares function

$$\begin{aligned} \|v\|_{\mathbf{W}}^2 = & \|z(t_0) - z_0\|_{\mathbf{P}_0}^2 \\ & + \sum_k \|\phi(x(t_k), U_k) - x(t_{k+1})\|_{\mathbf{Q}_k}^2 \\ & + \sum_j \|\mathbf{h}(z(t_j)) - \tilde{y}(t_j)\|_{\mathbf{R}_{y_j}}^2, \end{aligned} \quad (28.144)$$

where the optimization variable is the portion of the system trajectory contained in Z . The vector v is the concatenation of each of the vectors summed in the right-hand side of (28.144) and \mathbf{W} is the positive definite block diagonal matrix formed by the positive definite submatrices \mathbf{P}_0 , \mathbf{Q}_k , and \mathbf{R}_{y_j} . The quantity

$$\|v\|_{\mathbf{W}}^2 = v^T \mathbf{W}^{-1} v$$

is the squared Mahalanobis distance defined based on the matrix \mathbf{W} . The reduction of the cost function from (28.143) to (28.144) relies on the reasonable and standard assumptions that $v_u \sim N(0, \mathbf{Q}_d)$, $n_y \sim N(0, \mathbf{R}_y)$, and $x(t_0) \sim N(x_0, \mathbf{P}_0)$.

Efficient algorithms for real-time solution of (28.144) are presented, for example, in [28.47, 48]. In the SLAM literature, accurate estimation of the entire rover trajectory is of interest, as it is necessary

for the accurate estimation of the landmark feature map.

In GNSS-aided INS applications, the same techniques applied over short duration (e.g., 30 s) time windows containing recent measurements have utility for different objectives: detection of faulty or spoofed satellite signals, detection of multipath, detection of IMU faults, and INS initialization to name just a few. In such applications, the measurement sets Z , Y , and U are defined to contain a recent window of measurement, that is, all measurements for $t_k \in [t-M, t]$. In this approach, all measurements prior to $t-M$ have been deemed valid (or discarded) by prior processing so that $\hat{z}(t-M)$ is accurate with known error covariance. Realtime minimization of (28.144) provides the necessary state estimate, while also allowing comprehensive analysis of the residuals relative to the stated assumptions. Such methods have the ability to enhance the availability and continuity of solutions with high integrity. Related fault detection methods have a long history in the navigation literature [28.54], but mostly are applied at only the most recent point in time. Such present time approaches have the limitation that the effect of the fault on the navigation solution cannot be removed, if the fault is not detected at the time when it occurs. Alternatively, evaluation over a sliding window of recent measurements offers an enhanced ability to detect faults within the window as well as the capability to recompute a fault-free solution to the navigation problem when a fault is detected within the interior of the temporal window. This approach is applied in [28.55, 56] where compass-free initialization of a low cost INS is demonstrated. These methods have only recently become practical due to the increased performance and decreased cost of computation.

28.11 Looking Forward

This chapter began with the notion that the cost of sensors and computation was rapidly decreasing, whereas the capabilities of both were rapidly increasing. In addition, the number of independent GNSS's and their capabilities are rapidly increasing. GPS modernization is only one example of the increased number of signals and signal strength that is becoming available. These various factors taken together indicate a very bright future for the accuracy, integrity, and availability of navigation systems at cost points low enough for new and rapidly evolving commercial applications. Finally, there are several opportunities for aiding by non-GNSS sources: feature aiding via RADAR, light detection and ranging (LIDAR), or vi-

sion; digital communications-based ranging by DTV, digital radio, or cell phone (signals of opportunity (SOO)). In addition, some of these alternative signals, such as cell phones, are becoming reliable sources for communication of real-time GNSS differential corrections.

At the same time, recreational, commercial, sporting, and military users have all had a small sample of the interesting new applications possible from reliable and accurate real-time navigation solutions. Some of these applications – for example, autonomous vehicles, safety-augmentation systems, or manned safety-of-life systems – require high levels of integrity with predictive indicators of certain accuracy specification violations.

GNSS-aided inertial systems are the dominant technology with a bright future for such applications.

A current challenge for all existing approaches is the detection and mitigation of spoofing. This is

one challenge that may be addressable by integration of information from various sources, including GNS, IMU, vision, radar, signals-of-opportunity, and so on.

References

- 28.1 B.L. Stevens, F.L. Lewis: *Aircraft Control and Simulation* (Wiley, New York 1992)
- 28.2 Global Positioning System Wide Area Augmentation System (WAAS) Performance Standard (US Federal Aviation Administration, Washington DC 2008)
- 28.3 T. Walter, P. Enge: Weighted RAIM for precision approach, Proc. ION GPS 1995, Palm Springs (ION, Virginia 1995) pp. 1985–2004
- 28.4 S. Hewitson, J. Wang: GNSS receiver autonomous integrity monitoring (RAIM) performance analysis, GPS Solutions **10**(3), 155–170 (2006)
- 28.5 J.J. Gertler: Analytical redundancy methods in fault detection and isolation, Proc. IFAC/IMACS Symp. Fault Detect. Superv. Saf. Tech. Process. SAFEPROCESS'91, Baden-Baden (1991) pp. 9–21
- 28.6 E.Y. Chow, A.S. Willsky: Analytical redundancy and the design of robust failure detection systems, IEEE Trans. Autom. Contr. **29**, 603–614 (1984)
- 28.7 K.R. Britting: *Inertial Navigation Systems Analysis* (Wiley-Interscience, New York 1971)
- 28.8 J.A. Farrell: *Aided Navigation: GPS with High Rate Sensors* (McGraw-Hill, New York 2008)
- 28.9 J.L. Farrell: *Integrated Aircraft Navigation* (Academic, New York 1976)
- 28.10 C. Jekeli: *Inertial Navigation Systems with Geodetic Applications* (Walter de Gruyter, Berlin 2001)
- 28.11 J. Wendel: *Integrierte Navigationssysteme: Sensordatenfusion, GPS und Inertiale Navigation* (Oldenbourg Wissenschaftsverlag, Munich 2011), in German
- 28.12 P.D. Groves: *Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems* (Artech House, Norwood 2013)
- 28.13 M.D. Shuster: Survey of attitude representations, J. Astronaut. Sci. **41**(4), 439–517 (1993)
- 28.14 R.G. Brown, P.Y.C. Hwang: *Introduction to Random Signals and Applied Kalman Filtering*, 4th edn. (Wiley, New York 2012)
- 28.15 A.H. Jazwinski: *Stochastic Processes and Filtering Theory* (Academic, San Diego 1970)
- 28.16 P.S. Maybeck: *Stochastic Models, Estimation, and Control* (Academic, San Diego 1979)
- 28.17 D.W. Allan: Statistics of atomic frequency standard, Proc. IEEE **54**(2), 221–231 (1966)
- 28.18 P. Savage: Strapdown system algorithms. In: *Advances in Strapdown Inertial Systems*, AGARD Lecture Series, Vol. 133, ed. by G.T. Schmid (NATO Advisory Group for Aerospace Research and Development, Neuilly-Sur-Seine 1984), pp. 3.1–3.30
- 28.19 L.M. Silverman, H.E. Meadows: Controllability and observability in time-variable linear systems, SIAM J. Contr. **5**(1), 64–73 (1967)
- 28.20 I.Y. Bar-Itzhack, N. Bergman: Control theoretic approach to inertial navigation systems, J. Guid. **11**(3), 237–245 (1988)
- 28.21 F.M. Ham, R.G. Brown: Observability, eigenvalues, and Kalman filtering, IEEE Trans. Aerosp. Electron. Syst. **19**, 269–273 (1983)
- 28.22 D. Goshen-Meskin, I.Y. Bar-Itzhack: Observability analysis of piece-wise constant systems Part 1: Theory, IEEE Trans. Aerosp. Electron. Syst. **28**, 1056–1067 (1992)
- 28.23 A.B. Chatfield: *Fundamentals of High Accuracy Inertial Navigation* (AIAA, Reston 1997)
- 28.24 A. Lawrence: *Modern Inertial Technology: Navigation, Guidance, and Control*, 2nd edn. (Springer, New York 2001)
- 28.25 D.H. Titterton, J.L. Weston: *Strapdown Inertial Navigation Technology*, 2nd edn. (IEE, Stevenage 2004)
- 28.26 J. Wendel, G.F. Trommer: An efficient method for considering time correlated noise in GPS/INS integration, Proc. ION NTM 2004, San Diego (ION, Virginia 2004) pp. 903–911
- 28.27 W.A. Poor: A geometric description of wander azimuth frames, Navigation **36**(3), 303–318 (1989)
- 28.28 J.E. Bortz: A new mathematical formulation for strapdown inertial navigation, IEEE Trans. Aerosp. Electron. Syst. **7**(1), 61–66 (1971)
- 28.29 J.B. Kuipers: *Quaternions and Rotations Sequences* (Princeton Univ. Press, Princeton 1999)
- 28.30 B. Palais, R. Palais, S. Rodi: A disorienting look at Euler's theorem on the axis of a rotation, Am. Math. Mon. **116**(10), 892–209 (2009)
- 28.31 C. Broxmeyer: *Inertial Navigation Systems* (McGraw Hill, New York 1964)
- 28.32 W.R. Hamilton: On quaternions; or on a new system of imaginaries in algebra, The London, Edinburgh and Dublin Philos. Mag. J. Sci. **xxv**(3), 489–495 (1844)
- 28.33 W.A. Heiskanen, H. Moritz: Physical geodesy, Bull. Géod. **86**(1), 491–492 (1967), in French
- 28.34 W.H. Press, S.A. Teukolsky, W.T. Vetterling: *Numerical Recipes: The Art of Scientific Computing* (Cambridge Univ. Press, Cambridge 2007)
- 28.35 M. Schuler: Die Störung von Pendel und Kreiselapparaten durch die Beschleunigung des Fahrzeuges, Phys. Z. **24**(16), 344–350 (1923), in German
- 28.36 M. Grewal, A.P. Andrews: *Kalman Filtering: Theory and Practice Using Matlab* (Wiley, New York 2008)
- 28.37 S.M. Kay: *Fundamentals of Statistical Signal Processing, Estimation Theory* (Prentice Hall PTR, Upper Saddle River 1993)
- 28.38 D. Simon: *Optimal State Estimation: Kalman, H_∞ , and Nonlinear Approaches* (Wiley, Hoboken 2006)

- 28.39 R.E. Kalman: A new approach to linear filtering and prediction problems, *J. Basic Eng.* **82**(1), 35–45 (1960)
- 28.40 E. Ohlmeyer: Analysis of an ultra-tightly coupled GPS/INS system in jamming, *Proc. IEEE/ION PLANS 2006*, San Diego (ION, Virginia 2006) pp. 44–53
- 28.41 D. Gustafson, J. Dowdle, K. Flueckiger: A high anti-jam GPS-based navigator, *Proc. ION NTM 2000*, Anaheim (ION, Anaheim 2000) pp. 495–503
- 28.42 A. van Dierendonck, J. McGraw, R. Brown: Relationship between allan variances and Kalman filter parameters, *Proc. 16th Precise Time Time Interval (PTTI) Appl. Plan. Meet.* (1984) pp. 273–293
- 28.43 R.R. Hatch: The synergism of GPS code and carrier measurements, *Proc. 3rd Int. Geod. Symp. Satell. Doppler Position.*, Las Cruces (1982) pp. 1213–1232
- 28.44 R. Hatch: Instantaneous ambiguity resolution, *Proc. Int. Symp. Kinemat. Syst. Geod. Surv. Remote Sens.*, Banff, ed. by K.-P. Schwarz, G. Lachapelle (Springer, New York 1991) pp. 299–308
- 28.45 P.J.G. Teunissen: The least-squares ambiguity decorrelation adjustment: A method for fast GPS integer ambiguity estimation, *J. Geod.* **70**, 65–82 (1995)
- 28.46 P.J.G. Teunissen: GPS carrier phase ambiguity fixing concepts. In: *GPS for Geodesy*, ed. by A. Kleusberg, P. Teunissen (Springer, Berlin 1996) pp. 263–335
- 28.47 F. Dellaert, M. Kaess: Square root SAM: Simultaneous localization and mapping via square root information smoothing, *Int. J. Robot. Res.* **25**(12), 1181–1203 (2006)
- 28.48 M. Kaess, A. Ranganathan, F. Dellaert: iSAM: Incremental smoothing and mapping, *IEEE Trans. Robot.* **24**(6), 1365–1378 (2008)
- 28.49 Y.F. Jiang, Y.P. Lin: On the rotation vector differential equation, *IEEE Trans. Aerosp. Electron. Syst.* **27**(1), 181–183 (1991)
- 28.50 J.C. Fang, D.J. Wan: A fast initial alignment method for strapdown inertial navigation system on stationary base, *IEEE Trans. Aerosp. Electron. Syst.* **32**(4), 1501–1505 (1996)
- 28.51 S. Thrun: *Probabilistic Robotics* (MIT Press, Cambridge 2005)
- 28.52 B. Triggs, P. McLauchlan, R. Hartley, A. Fitzgibbon: Bundle adjustment – A modern synthesis, *vision algorithms: Theory and practice* **1883**, 298–372 (2000)
- 28.53 A. Vu, J.A. Farrell, M. Barth: Centimeter-accuracy smoothed vehicle trajectory estimation, *IEEE Intell. Transp. Syst. Mag.* **5**(4), 121–135 (2013)
- 28.54 J.C. Wilcox: Competitive evaluation of failure detection algorithms for strapdown redundant inertial instruments, *J. Spacecr.* **11**(7), 525–530 (1974)
- 28.55 Y. Chen, D. Zheng, P. Miller, J.A. Farrell: Underwater vehicle near real time state estimation, *Proc. IEEE Int. Conf. Contr. Appl.*, Hyderabad (2013) pp. 545–550
- 28.56 A. Ramanandan, J.A. Farrell, A. Chen: A near-real time nonlinear state estimation approach with application to initialization of navigation systems, *Proc. 50th IEEE Conf. Decis. Contr. Eur. Contr. Conf.*, Orlando (2011) pp. 3184–3191

Land and Maritime Applications

Allison Kealy, Terry Moore

This chapter presents an overview of applications of global navigation satellite systems (GNSS) relevant to the land and maritime environments. It focuses on the positioning performance requirements, technologies, developments, and trends for current and projected growth areas for GNSS in the land, rail, and maritime transport sectors. Representative applications including: personal navigation, location-based services, maritime, and land-based intelligent transport systems, railway logistics, and maritime operations are showcased as they encompass overlapping and related tasks such as fleet and asset monitoring, cooperative mobility, autonomous and precise navigation, vehicle and machinery control, and others.

29.1	Land-Based Applications of GNSS	842
29.1.1	Personal Devices.....	843
29.1.2	Location-Based Services.....	845
29.1.3	Positioning Technologies and Techniques for PN and LBS.....	846
29.1.4	Intelligent Transport Systems.....	853
29.2	Rail Applications	856
29.2.1	Signaling and Train Control.....	857
29.2.2	Freight and Fleet Management.....	862
29.2.3	Passenger Information Systems.....	863
29.3	Maritime Applications	863
29.3.1	GNSS Performance Requirements for Maritime Applications.....	864
29.3.2	Maritime Navigation.....	867
29.3.3	eLoran.....	869
29.3.4	Automatic Identification System.....	869
29.3.5	Shipping Container Tracking.....	872
29.4	Outlook	873
	References	873

Applications across the land and maritime domains have traditionally been classified as high end or low end, where these two classes are distinguished primarily by their accuracy requirements. High-end applications such as land and hydrographic surveying (Chap. 35) or geodesy and mapping (Chap. 36) have positioning accuracy requirements at the centimeter or subcentimeter level and low-end applications such as personal navigation (PN) and other mass-market (consumer-level) applications are satisfied with accuracies at the meter, submeter or even tens of meters level. The pervasiveness and convenience of global navigation satellite systems (GNSS), combined with an increasing recognition of the value of positioning accuracy, has created an unprecedented demand for better levels of GNSS performance, blurring the distinction between high- and low-end applications. Significantly, with the increasing ubiquity of GNSS and mounting concerns surrounding the vulnerability of GNSS, applications are now being classified as safety-liability-critical or nonsafety-liability-critical, where positioning needs have extended well beyond that of just accuracy. Parameters such as

positioning integrity, reliability, continuity, and availability have emerged as the key performance metrics underpinning the technologies and techniques selected for individual applications. In this chapter, definitions for these parameters are adapted from [29.1]. Accuracy is defined as the degree of conformance of an estimated or measured position at a given time to a reference value. Reliability refers to the ability of the system to detect blunders in the measurements and to estimate the effects of undetected blunders on the position solution. Integrity refers to the probability (per operation or per unit of time) that the system generates an unacceptable error also without providing a timely and valid warning to users that the system must not be used for the intended operation. Availability is the percentage of time the system is able to provide solutions within specified accuracy, reliability, and integrity thresholds. Continuity refers to the probability that the system will stop providing position outputs of the specified quality during a given operation or time interval.

According to the GNSS market monitoring report published by the European GNSS Agency

(GSA) [29.2], by 2019 there will be on average, one GNSS receiver in use for every person on the planet. This trend is driven by the enhanced performance capabilities and competitive pricing of smartphones and tablets. With the increasing pervasiveness of mobile, location-enabled devices, new applications are continually being introduced as both consumers and industry sectors recognize the efficiencies that can be gained from knowing where things are. It would be unfeasible to provide a complete study of all of these applications in a single chapter. This chapter therefore focuses on current and projected growth areas in land-based and maritime applications of GNSS. It is divided into three sections.

In the first section, land-transport applications are presented. It introduces the state of the art in PN systems and the trends in location-based services (LBS) brought about by the increasing ubiquity of mobile devices including smartphones and tablets. Positioning technologies and techniques that address some of the vulnerabilities of GNSS for applications that leverage the location-aware capabilities of current generation

smartphones are also presented. Following on from this, intelligent transport systems (ITS) are presented as a class of applications that combine both consumer- and enterprise-level LBS applications and cut across safety- and nonsafety-critical requirements of positioning performance.

Demands for positioning integrity have seen the rail sector lag significantly behind that of the land and maritime sectors in terms of the adoption of GNSS. In the second section, rail-transport applications are therefore treated as a distinct class of land applications. In this section, a range of safety- and nonsafety critical applications that mirror and complement the objectives for ITS in the road transport sector are described. Future GNSS is offering significant potential for integrity enhancement and is the subject of a number of investigative studies for the rail transport sector. The outcomes and recommendations of some of these studies are presented in this section.

In the third section, maritime applications of GNSS that support and contribute to the development of maritime ITS (MITS) are presented.

29.1 Land-Based Applications of GNSS

According to [29.2], the trend is for LBS and road segments to dominate GNSS cumulative revenue, with a combined total of more than 91% (Fig. 29.1). This is driven by the increasing capability for smartphones and tablets to replace dedicated, nomadic personal navigation GNSS devices (PND) such as in-car navigation systems, as well as increased uptake in location-aware applications and data services.

The expansion of GNSS use in the land transport sector will continue into the future given the expected growth and adoption of personal navigation applications (PNAs), LBS, and ITS. Offering the potential to signif-

icantly improve the safety and efficiency of road networks, these applications extend well beyond the traditional domains of vehicle navigation, commercial fleet management, public transport monitoring, passenger information, and emergency vehicle location and dispatch. GNSS now underpins the capabilities of safety-critical applications that operate under the ITS concept. A study undertaken by the Royal Academy of Engineering [29.3] provides a representative list of land, rail, and maritime GNSS applications. Table 29.1 shows the breadth of land transport applications relying on GNSS and the diverse positioning accuracy levels required.

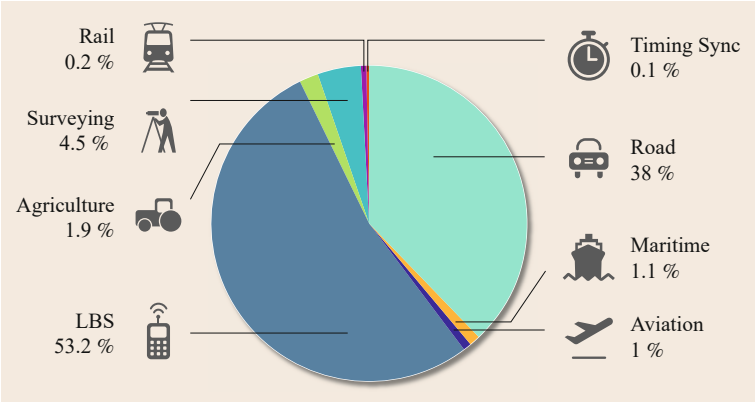


Fig. 29.1 Cumulative core revenue by segment (cumulated revenue 2013–2023) (after [29.2], courtesy of European GNSS Agency, 2015)

Table 29.1 Accuracy requirements for land transport applications. Road level accuracy: typically meter level positioning at 1 Hz. Lane level accuracy: typically submeter level positioning at 1 Hz. Where-in-lane level accuracy: typically decimeter–centimeter level positioning at 1 Hz (after [29.3])

Application	Accuracy
In-car navigation	Road level
Fleet management	Road level
Urban traffic control	Lane level
Emergency calls	Road level
Dynamic route guidance	Road level
Selective vehicle priority	Road level
Collision avoidance	Where-in-lane level
Automated highway	Where-in-lane level
Road pricing	Lane level
Intelligent speed assistance	Where-in-lane level
Lane control	Where-in-lane level
Stolen vehicle recovery	Road level
Restraint deployment	Where-in-lane level
Trip travel information	Road level

In this section, the GNSS positioning techniques underpinning land transport applications are presented and the use of augmentation infrastructure and assistance data to improve the accuracy and availability of the position solution are described. As these applications mature, and their use become more widespread, a range of topics are emerging as relevant and significant. Of particular focus to this chapter are the techniques, sensors, and signals that improve the robustness of the positioning solution for safety critical applications. This section therefore addresses important topics related to the fusion of GNSS and non-GNSS positioning technologies. While we acknowledge the need to take into account other factors that impact on the utility of location-aware devices including privacy, user behavior, context awareness, and so on, they are considered outside of the scope of this chapter.

29.1.1 Personal Devices

Personal devices including not only smartphones and tablets but also specific equipment such as tracking devices, digital cameras, portable computers, and fitness gear support a plethora of location-based applications. There are almost three billion mobile applications currently in use that rely on positioning information. Of these, PN is perhaps the most mature, with in-vehicle navigation systems the most popular example.

PN is defined as a means of determining an individual's location (positioning), selecting and providing guidance on the required route and mode of transport to a desired destination in *any* environment (outdoor or indoor), using geographic or spatial information, and information on location-based phenomena and services, [29.4]. These systems typically provide guidance information that assist a user in navigating from a point of origin to a destination. PN capabilities have resulted from the convergence of mobile computing, spatial/mapping information, and positioning technologies, with the first-generation PN devices (PNDs), dedicated units providing simple Global Positioning System (GPS) point positions to pinpoint the user's location, overlaid onto a base map.

Until recently, the proliferation of in-vehicle navigation systems established PNDs as the largest consumer market for GNSS-enabled devices. However, these capabilities have now been integrated into applications that enable any location aware smartphone to operate as a PND. The inherent synergy between location and mobile communication facilitated in smartphones has enabled the *services* characteristic of PN. Navigation applications that use a variety of positioning hardware built into smartphone, and the communication capabilities of the device, can access real-time traffic updates, weather information, and other relevant spatial and temporal information to improve the navigation instructions provided (Fig. 29.2).



Fig. 29.2 (a) TomTom in-car navigation system as a personal navigation application for a smartphone and (b) TomTom in-car navigation system as a standalone PN device (courtesy of TomTom International BV)

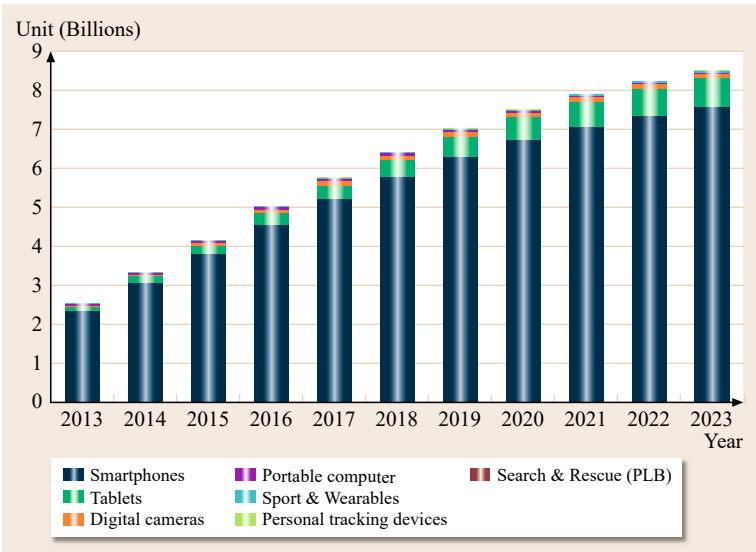


Fig. 29.3 GNSS devices in use by application – LBS segment (after [29.2], courtesy of European GNSS Agency, 2015)

Table 29.2 Classification of LBS by applications (after [29.5])

Location-based services	Applications	Required quality of service (QOS)
Information/directory services	<ul style="list-style-type: none">Dynamic yellow pages that automatically informs consumer of location of nearest hospitals, restaurants, shopping malls and theatre, and ATMNearest parking lot, drug store, or gas station	Location accuracy of a tens of meters Response time of few seconds Need for high reliability (98–99%)
Tracking and navigation services	<ul style="list-style-type: none">Tracking of children, locating lost petsLocating friends in a particular areaTracking stolen vehicles, asset trackingDynamic navigational guidanceVoice-enabled route description	Location accuracy of few meters Response time of few seconds Need for very high reliability (Goal should be 100%)
Emergency services	<ul style="list-style-type: none">Roadside assistanceSearch and rescue missionsPolice and fire responseEmergency medical ambulance, E911	Location accuracy of a tens of meter Response time of few seconds or less Need for very high reliability (Goal should be 100%)
Location-based advertising	<ul style="list-style-type: none">Wireless coupon presentation, targeted and customized adsMarketing promotions and alertsCustomer notification and identification in the neighborhood store	Location accuracy of few meters Response time of a minute Need for high reliability (98–99%)

The distinctions between PNAs that run on a smartphone and dedicated PNDs are often argued in terms of cost and usability – applications typically cost less and operate on a single mobile device and PNDs are more power efficient and simple to use. However, as the functionality and use of smartphones increases, the decline evident in the standalone PNDs market will continue. Figure 29.3 shows the number of GNSS-based devices predicted to be in use through to 2023. While niche personal devices such as personal trackers and wearable devices are expected to gain prominence, smartphones and tablets will con-

tinue to dominate the landscape for personal devices. In fact, smartphones or devices that emulate smartphone functionality (messaging, Internet search, social apps, navigation, video and picture capture, etc.) such as Google Glass and Apple iWatch support an increasing range of location aware applications and services, including [29.2]:

- *Geo marketing and advertising:* Consumer preferences are combined with positioning data to provide personalized offers to potential customers and create market opportunities for retailers.

- **Safety and emergency:** GNSS, in combination with network-based methods, provides accurate emergency caller location.
- **Enterprise applications:** Mobile workforce management and tracking solutions are implemented by companies to improve productivity.
- **Sports:** GNSS enables monitoring of users' performance through a variety of fitness applications, such as step counters and personal trainers.
- **Games and augmented reality:** Positioning and virtual information are combined to entertain the user and improve everyday life.
- **Social networking:** Friend locators provided by dedicated apps or embedded in social networks use GNSS to help keep in touch and share travel information.

29.1.2 Location-Based Services

LBS refer to an increasing range of mobile computing applications that provide information and functionality to users based on their geographical location [29.7]. According to [29.8], all location-based user needs can be characterized into five fundamental mobile actions: identifying the location of the user with respect to another person or place – *locating*; searching for persons, objects, or events – *searching*; requesting directions to a location – *navigating*; requesting specific attributes of a location – *identifying*; looking for events at or near a certain location – *checking*. A simple classification of LBS applications is presented in Table 29.2.

In many cases, it is the services aspect that differentiates LBS from the autonomous devices and applications. For example, a sports watch used by an athlete to monitor speed or distance, as well as health monitoring indicators such as heart rate and blood pressure, differs from the services aspect of an alerting system that automatically requests assistance from emergency services based on tracking information received from a location-based medical device used by an elderly person.

LBS can be classified into two types: push and pull. In a push type of service, the user receives information from the service provider without requesting it at that instant, although the user may have originally subscribed to the service at an earlier time. For example, send an SMS message advertising a sale at a nearby shopping mall or an accident up ahead. In a pull type, the user has to actively request for information. For example, where is the nearest restaurant or train schedules. More complex LBS that are emerging can *push* information in response to user preferences, based on their location, making the experience more personalized or *context-aware*. For example, *tell me when I'm near the library as I have a book to return*. Other types of LBS based on service distinctions can be found in Table 29.3.

The generalized architecture of LBS is shown in Fig. 29.4. At the core of these systems is the positioning information used to describe the location of the mobile handset. In the following sections, the typical positioning technologies used in current generation LBS are fully described.

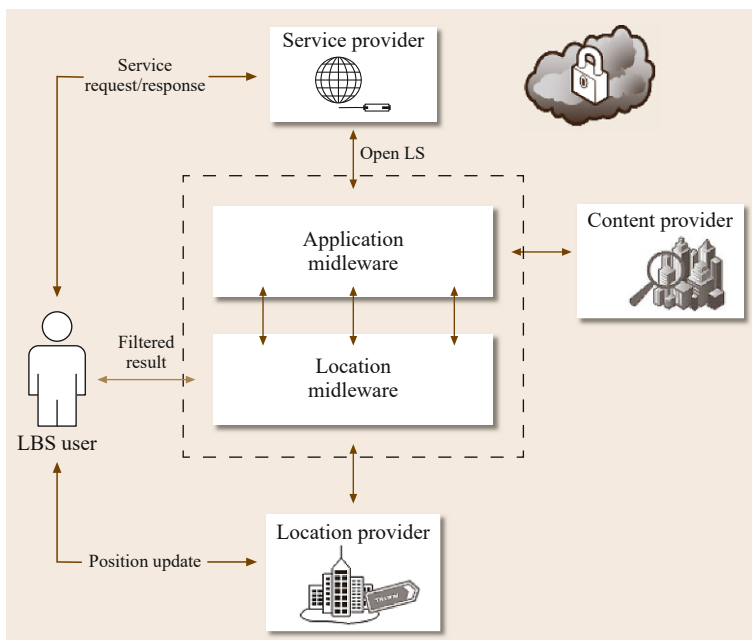


Fig. 29.4 Components of an LBS (after [29.6])

Table 29.3 Classification of LBS by applications (after [29.5])

Types of LBS	Characteristics
Person-oriented LBS	<ul style="list-style-type: none">● Consists of applications where a service is user based● User usually controls how location information is collected and utilized
Device-oriented LBS	<ul style="list-style-type: none">● Applications are external to user● Person or the device located is not controlling the service
Push- versus pull-based applications	<ul style="list-style-type: none">● Push based: information delivered to the mobile terminal (end user) automatically when certain event occurs● Pull based: mobile terminal (end user) initiates the request
Direct versus indirect profile	<ul style="list-style-type: none">● Based on how the user profile is collected: directly from the user during the set up phase, by tracking the user's behavior pattern or from third parties● Security and privacy issues become critical to maintain user trust and to avoid fraudulent activities
Availability of profile information	<ul style="list-style-type: none">● Profile information requested on the fly or already available to the LBS
Mobility and interaction	<ul style="list-style-type: none">● Range of mobility scenarios exist based on combinations of mobility of users and network components● The level and type of interactions depend on the mobility scenario
State of interaction	<ul style="list-style-type: none">● Stateless interaction: each request is an independent transaction unrelated to previous request● Stateful interaction: the LBS preserves the state across service requests (beneficial to for forecasting future transactions, requests, and behavior)
Static versus dynamic information source	<ul style="list-style-type: none">● Static: data about historical buildings and landmarks, places of attraction, hotels and restaurants, and maps● Dynamic: information that changes with time (weather, traffic, and road conditions)
Source of location information	<ul style="list-style-type: none">● Location information provided by the user or the network infrastructure or by a third party
Accuracy of location information	<ul style="list-style-type: none">● Depending on the positioning technology used in the network infrastructure, different accuracy for localization request of mobile terminals result

29.1.3 Positioning Technologies and Techniques for PN and LBS

Since 2011, the GNSS positioning hardware integrated into smartphones has been based around multiple GNSS constellations – combining signals from both the GPS and GLONASS constellations. More recently, receiver chipsets that integrate signals from combinations of or all constellations have become available or an under evaluation (e.g., Broadcom’s BCM47531, a GNSS chip that generates positioning data from five satellite constellations simultaneously (GPS, GLONASS, QZSS, SBAS, and BeiDou). Figure 29.5 shows the percentage of available receivers capable of tracking signals from one GNSS (i.e., GPS only), two GNSS (i.e., GPS + Galileo, GPS + GLONASS, GPS + BeiDou), three GNSS (i.e., GPS + Galileo + GLONASS, GPS + Galileo + BeiDou, GPS + GLONASS + BeiDou), or tracking signals from all constellations at the same time. It can be concluded that almost 60% of all available receivers, chipsets, and modules are supporting a minimum of two constellations.

These receivers typically deliver a solution commensurate with that of the GPS standard positioning

service (SPS), and are among the lowest cost receivers on the market (consumer grade). Currently, SPS capabilities are provided by the coarse acquisition (C/A), pseudorandom noise code transmitted on the L1 frequency. This positioning accuracy is formally specified (for GPS) as better than 9 m, 95% horizontal error and 15 m, 95% vertical error. More recently, results of 3.286 m horizontal and 6.301 m vertical were recorded for the first quarter of 2014, by regular GPS SPS assessments undertaken by the US Federal Aviation Association at sites across the United States [29.9].

Over the last decade, the satellite positioning technology for an increasing range of location-aware devices has matured, offering much better performance in terms of sensitivity, power consumption, size, and cost than was previously possible. These receivers offer more channels, capable of tracking signals from GNSS constellations as well as regional satellite-based augmentation systems (SBAS) including the European Geostationary Navigation Overlay Service (EGNOS) which covers Western Europe and beyond, the Wide Area Augmentation System (WAAS) covering North and Central America, the multifunctional satellite augmentation system (MSAS) covering Japan and East

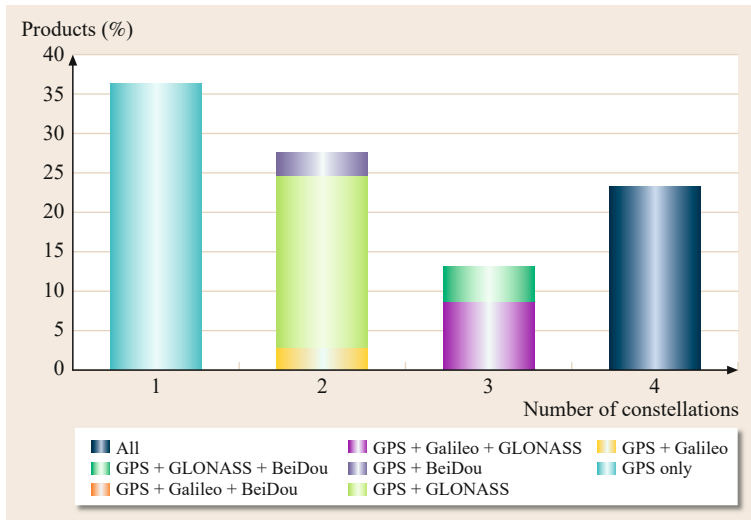


Fig. 29.5 Supported constellations by receivers – LBS segment (after [29.2], courtesy of European GNSS Agency)

Asia, the GPS and **GEO** Augmented Navigation System (**GAGAN**) to cover the Indian subcontinent, and the system for differential corrections and monitoring (**SDCM**) as a component of GLONASS.

Figure 29.6 shows the u-blox EVA-7M chipset, typical of the hardware integrated into mobile devices. Table 29.4 provides a summary of the corresponding technical specifications for this receiver.

An increase in the number of signals being tracked is aimed at enhancing both the positioning availability and accuracy. In GNSS difficult operating environments, such as buildings, tunnels, and urban streetscapes, satellite visibility is completely or partially obscured or multipathing effects can dramatically degrade the measurement quality. As these conditions are increasingly characteristic of the environments in which PNAs are most commonly used, modern PNAs adopt a range of techniques to improve positioning accuracy and availability, these include the use of high sensitivity GNSS receivers (HSGNSS), Assisted GNSS

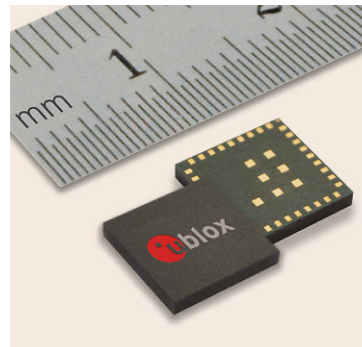


Fig. 29.6 u-blox PN chipset (courtesy of u-blox)

(**A-GNSS**), SBAS corrections, and sensor/measurement fusion techniques.

High Sensitivity GNSS

HSGNSS receivers have become the preferred choice for PND and LBS because of their ability to acquire and track the weak GNSS signals available in degraded

Table 29.4 Technical specification for the u-blox EVA-7M chipset (after [29.10]). **CEP** (circular error probability) is defined as the radius of a circle centered on the true value that contains 50% of the actual measurements

Receiver type	56-channel u-blox 7 engine, GPS/QZSS L1 C/A, GLONASS L1 frequency division multiple access (FDMA), SBAS (WAAS, EGNOS, MSAS)	
Navigation	update rate up to 10 Hz	
Accuracy (GPS/GLONASS)	Position	2.5 m/4.0 m CEP
	SBAS	2.0 m/n.a. CEP
Acquisition (GPS/GLONASS)	Cold starts	30 s/32 s
	Aided starts	5 s/n.a.
	Reacquisition	1 s/3 s
Sensitivity (GPS/GLONASS)	Tracking	−160/−158 dBm
	Cold starts	−147/−139 dBm
	Warm starts	−148/−145 dBm

signal environments. For example, GPS signals transmitted some 20 000 km above the Earth are already very weak, typically -160 dBW when they arrive at a receiver on the surface of the Earth. In outdoor environments, GPS signals are around the -155 dBW level. According to [29.11], attenuation of the signal by trees, buildings, etc., can reach values of about 5 dB in cars, up to 20 dB in buildings and more than 25 dB in subterranean garages. Standard GNSS receivers are not sensitive enough to track these signals with low dBW values and are therefore unable to work in these environments. A high sensitivity receiver can acquire, track, and compute a position from a weak signal that is 1/1000 the strength of a typical outdoor signal.

To compute a position, a GNSS receiver has to perform two operations; the first is termed *acquisition* and the second *tracking*. A minimum of four satellites need be tracked in order for a three-dimensional position solution to be obtained. In the acquisition phase, the receiver first assumes that a satellite is visible and allocates a channel to this satellite. There are two search unknowns. One is the exact frequency of each satellite carrier – which needs to be tuned in response to the inherent uncertainty in the receiver reference oscillators and the Doppler uncertainty due to inaccuracies in the receiver location and velocity. The other is the alignment (correlation) of the received signal with a locally generated replica of the code. To acquire the signal, a GNSS receiver must therefore search the entire space of possible frequency offsets and code delays.

Conventional GNSS receivers integrate the code delays for each frequency offset for 1 ms, which is the duration of a complete C/A code cycle. To improve the acquisition of weak signals in GNSS difficult environments, high sensitivity receivers increase the integration time and the number of correlators available. Commercially available GNSS receivers typically have only 2–4 correlators per channel while high-sensitivity receivers, such as the u-blox positioning engine and SiRFstar chipsets from Qualcomm, have over two million correlators. By significantly increasing the number of correlators in the receiver, many frequency/delay searches can be conducted in parallel, with more correlations and integrations possible than with a standard receiver in the same amount of time. The end result is an increase in the sensitivity of the receiver, without increasing the acquisition time. To see this, let n denote the number of correlators available for each satellite, T denote the integration time, A denote the acquisition time, and B denote the number of code delay and Doppler bins. These quantities are related by

$$A = \frac{BT}{n}. \quad (29.1)$$

It is clear from (29.1) that, for a fixed acquisition time A and number B of search bins, increasing the number n of correlators allows the integration time T to be increased commensurately. The benefit of this is that increasing T increases the output signal-to-noise ratio (SNR) of the correlators therefore allowing weaker input signals to be detected. It is wellknown that, as T increases, the power P at the output of the correlator is approximately proportional to T , i. e., $P = P_0 GT$ where the quantity P_0 is the input power and G is the power gain per second of integration time. After substituting $T = P/(P_0 G)$, we can then rearrange (29.1) to find an expression for the power at the correlator output

$$P = \frac{AP_0 Gn}{B}. \quad (29.2)$$

A conventional receiver has $B = 120 \cdot 10^3$ bins to search in $A = 60$ s using $n = 4$ correlators. If the signal arrives at the receiver with $P_0 = 10^{-16}$ W, then the correlator output power is

$$\begin{aligned} P &= 60 \cdot 10^{-16} \times 4G / 120 \cdot 10^3 \\ &= 2G \cdot 10^{-19} \text{ W}. \end{aligned} \quad (29.3)$$

A HSGPS receiver with $n = 1000$ correlators achieves the same output power for a received signal with power

$$\begin{aligned} P_0 &= \frac{BP}{AGn} \\ &= 120 \cdot 10^3 \times 2G \cdot 10^{-19} / (60G \cdot 10^3) \\ &= 4 \cdot 10^{-19} \text{ W} \approx -183.8 \text{ dBW}. \end{aligned} \quad (29.4)$$

Thus, increasing the number of correlators greatly increases the sensitivity of the receiver. Note that for integration times greater than 20 ms noncoherent integration must be performed.

If the attenuation is too large, the acquisition time can still be too long. This problem can only be solved if external information is available to reduce the search space for individual correlators. Once at least four satellites are being tracked, and their ephemeris is decoded, a receiver can perform the position, velocity, and time calculations. If it loses lock, then acquisition must be repeated, although the search space is now considerably smaller and therefore high sensitivities can be accommodated for tracking with little overhead in terms of computation and hardware.

Assisted GNSS

Assisted GPS (A-GPS) or assisted GNSS (A-GNSS) is closely linked to HSGNSS but is particularly relevant to LBS, as it leverages the communication capabilities of mobile devices to enable more rapid acquisition

of the satellite signals and in particular weaker signals, significantly reducing the *time to first fix* (TTFF) for the GNSS receiver. A-GPS first gained prominence as the positioning technique underpinning the United States Federal Communications Commission's (FCC) Enhanced 911 (E911) program directive, which mandated that the accurate location of mobile phones had to be made available to emergency services. It is now a standard configuration for GNSS chipsets used for PN and LBS, offering more robust positioning capabilities in GNSS difficult environments.

A-GNSS capabilities are directed at the positioning hardware integrated into mobile handsets – the C/A code on the L1 frequency. As mentioned in the previous section, before the GNSS receiver can compute its own position, it needs to be able to acquire and track the GNSS signals. To achieve this, the receiver needs to know what satellites are visible and then be able to decode orbital information for these satellites from the navigation message. In particular, the satellite clock parameters (in subframe 1 of the navigation message) and ephemerides (in subframes 2 and 3) are essential to computing the position solution. The satellite receiver uses the satellite clock corrections to determine the precise time that the signal was sent, and the orbital parameters are used to calculate the satellite position in space at that time. If the signal level is below about -173dBW , an unassisted receiver may not be able to achieve a valid fix at all because the signal is too weak for the receiver to decode the navigation data message. This situation is exacerbated by the inferior quality of

the antenna available in smartphones compared to other standalone PNDs.

In GNSS difficult environments, the mobile phone network is used to assist the receiver in the mobile device to overcome the problems associated with TTFF and the low signal levels that are encountered. A-GNSS provides information to the receiver as to what satellites (frequencies) it can expect to see and then provides assistance in the form of the positions of the satellites. In this way, the TTFF is reduced from the order of 1 min to a few seconds [29.12]. Figure 29.7 shows the typical A-GNSS architecture.

There are two techniques used to provide A-GNSS assistance:

1. *Mobile station based (MSB)*: In MSB mode, the mobile station performs the calculation. The A-GNSS device receives ephemeris, reference location, reference time, and other optional assistance data from the A-GNSS server. The A-GNSS device can then use this information to receive signals from visible satellites and calculate the position. The device can report the position, velocity, and time solution back to the A-GNSS server if required.
2. *Mobile station assisted (MSA)*: In MSA mode, the A-GNSS server performs the calculation. The A-GNSS capable device receives acquisition assistance, reference time, and other optional assistance data from a mobile service provider. Using this information, it acquires the signals of the visible satellites, makes the pseudorange measurements,

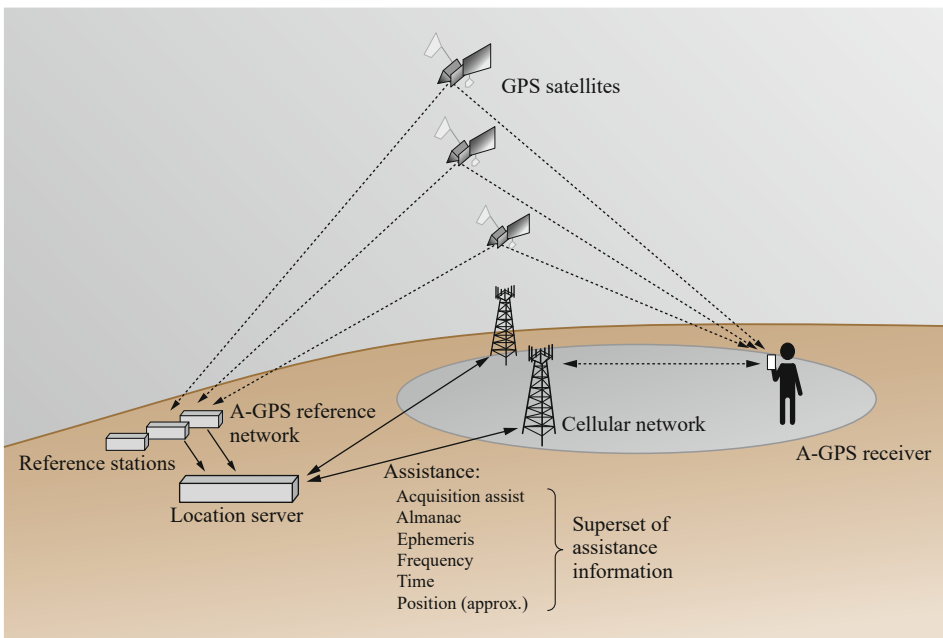


Fig. 29.7 Representative A-GNSS Architecture. (after [29.12] with permission)

and then sends these measurements to the A-GNSS server. The mobile service provider continuously logs GNSS information (mainly the almanac) from the GNSS satellites using a A-GNSS server in its system. With the help of the above data (the data received from the mobile device and the data already present in A-GNSS server) the A-GNSS server calculates the position and optionally sends it back to the A-GNSS device.

Comprehensive technical details on A-GNSS implementations can be found in [29.12].

Space-Based Augmentation Systems

Full details on space-based augmentation systems (SBAS) can be found in Chap. 12. Suffice to state here that SBAS capabilities augment the primary GNSS constellations providing ranging, integrity, and correction information. Primarily motivated by the integrity needs of the aviation sector, SBAS can significantly contribute to improving the accuracy, availability, and integrity of safety and liability critical applications such as tracking of the elderly and other vulnerable personnel, or collision avoidance systems.

The augmentation information provided by SBAS include corrections and integrity for satellite position errors, satellite clock errors, and ionospheric delays. These corrections are broadcast via geostationary satellites to suitably equipped satellite receivers. The accuracy improvements are at the levels of 1–3 m, 95% and are significant to many of the land transport applications in Table 29.1. While the integrity needs of the land transport sector are yet to be defined, their relevance and significance particularly in GNSS difficult environments have already been recognized.

Alternatives to GNSS

The positioning requirements of LBS have driven the need for GNSS-like or enhanced GNSS performance in environments that completely or partially obscure GNSS signals. Smartphones attempt to address this problem by using available signals in the phone to maintain the availability of the positioning solution in indoor and outdoor environments, switching between GNSS and other positioning modes such as WiFi and cell-based systems. In most instances, these positioning capabilities are more than adequate for the current generation of PN applications. However, for other navigation and tracking applications, other sensors and signals are offering more robust capabilities albeit with additional cost, infrastructure, and computational overheads.

In this section, we acknowledge that the major focus for non-GNSS positioning capabilities (for LBS) centers around indoor and urban environments. First, we

will look at the positioning technologies integrated into mobile phone handsets and second, a review of other signals and sensors being pursued as robust augmentations and alternatives to GNSS is provided. Figure 29.8 presents an overview of the accuracy and coverage characteristics of alternative/indoor positioning technologies [29.13].

Positioning with Cellular Networks. The full capabilities of LBS are delivered when the cellular networks and mobile handset work together to locate the user and then transfer the position data either upon request or continuously. Commercial examples range from low-accuracy positioning methods based on cell identity to high-accuracy methods combining wireless network information and satellite positioning (A-GNSS). The techniques can be classified as user centric – the position computation is performed by the user’s device, or network centric – the user’s position is determined by the tower or a hybrid solution. Table 29.5 provides a summary of the positioning techniques used by smartphones. Positioning methods include:

- The cell-of-origin (COO) method determines (at the mobile) which is the closest tower delivering the signal to the user. While this is an inexpensive method, the accuracy is limited to the density of towers or the size of the cell, which may range from 10–500 m for an indoor micro cell to an outdoor macro cell reaching several kilometers [29.14].
- Radiolocation techniques (similar to GNSS) uses characteristics of the radio signal that travels between the mobile use and the cellular towers to derive the handset position. By measuring the distance to at least three towers, positions of mobile users can be computed.
 - Time of arrival (TOA) is a network-centric approach that measures the time it takes for radio signals to arrive at multiple points.
 - Received signal strength (RSS) determines the distance from the measured signal strength. Its performance is significantly worse than any other cellular positioning technique as it is influenced by antenna direction and multipath effects.
 - Enhanced observed time difference (E-OTD) measures the difference between the time of arrival of the cellular signal at the handset and at a nearby fixed receiver. Time differences from at least three noncollinear towers are required to compute a position. The positioning accuracy of E-OTD is about 100–125 m.
- The location (multipath) pattern matching method uses the multipath signature in the vicinity of the mobile user to find its location. The user’s termi-

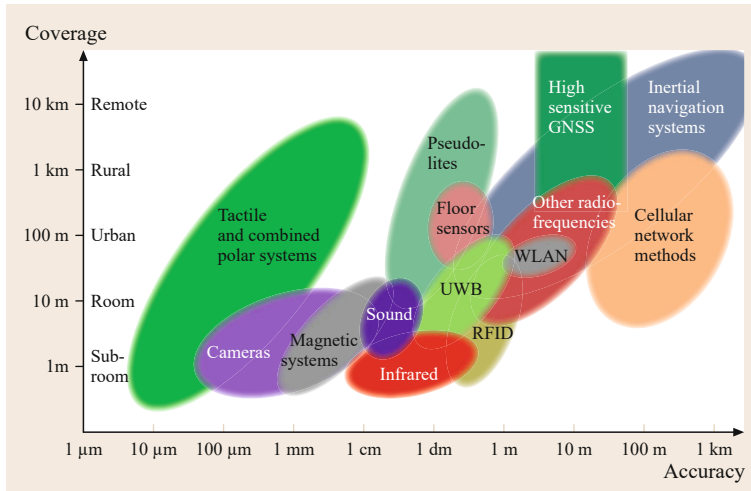


Fig. 29.8 Relationship between coverage and accuracy for a range of indoor positioning technologies. (after [29.13])

Table 29.5 Position determination techniques used in smartphones

Technique	Range/operational availability	Accuracy	Services and content
COO	Globally (in principle)	250 m–35 km	Traffic information, information services
E-OTD	Globally (in principle)	100–500 m	Location-based billing, information services
TOA	Globally (significant capital investment)	100–500 m	Location-based billing, information services
Fingerprint	In urban areas (significant capital investment)	< 150 m	Fleet management, advertisement, routing
GNSS	Globally	5–10 m	Positioning and navigation
A-GNSS	Globally (after significant capital investment)	5–10 m	Positioning and navigation
WLAN	Local (dependent on existing infrastructure)	1–10 m	Information services

nal creates a signal pattern (phase and amplitude characteristics) which can be measured and stored in a database. The cell tower receives a multipath signal from the mobile terminal and compares its signature with the multipath location database, which defines locations by their unique multipath characteristics.

Positioning Using Wireless Local Area Networks (WLAN). The widespread availability of WLAN (WiFi) networks present additional signals that can be used for positioning. The classic approach is trilateration, using ranges derived from the RSS measurements. However, the WiFi signals themselves are highly sensitive to interference and multipath caused by the operating environment (e.g., walls, people, equipment) and can result in incorrect positioning of the user. The method of fingerprinting (similar to pattern matching in cellular networks) is proving to be more successful but requires an additional workload to create the database containing the recorded signal strength data from various access points at known points spread across the environment in which positioning information is required. The Ekahau real-time location system (RTLS) uses a general propagation model for the WiFi signals, and parameterizes this model using a number of test measurements. The mobile user then sends its mea-

surements of signal strengths to all surrounding access points to the positioning engine, which in turn calculates the position by solving a maximum likelihood problem.

Ultrawideband Positioning. It is expected that smartphone positioning will, out of necessity, encompass a hybrid solution, based on the fusion of multiple sensors. It can also be expected that additional sensors and signals will find their way into smartphones – where, at this stage their role has been toward higher end (enterprise or commercial) applications given their additional cost and infrastructure requirements. Table 29.6 provides a summary of the characteristics of a range of alternative positioning technologies. These technologies are however, now routinely integrated into a range of the so-called real-time location systems (RTLS) where robust positioning performance in indoor or GNSS difficult environments negates the additional infrastructure costs associated with these technologies. Ultrawideband (UWB) is one example of this.

The high bandwidth and accurate pulse timing offered by UWB signals make them ideal for positioning given their high multipath resistance, penetration, and accurate ranging capabilities. Commercial UWB positioning systems such as that offered by

Table 29.6 Summary of the characteristics of alternative positioning technologies used in GNSS difficult environments

Technology	Typical accuracy	Typical coverage (m)	Measuring principle	Application
Cameras	0.1 mm–dm	1–10	Angles from images	Metrology, robot navigation
Infrared	cm–m	1–5	Thermal imaging, active beacons	People detection, tracking
Tactile & polar	μm–mm	3–2000	Mechanical, interferometry	Automotive, metrology
Sound	cm	2–10	Distance from time of arrival	Hospitals, tracking
RFiD	dm–m	1–50	Proximity detection and fingerprinting	Pedestrian, navigation
Ultrawideband	cm	1–50	Body reflection, time of arrival	Pedestrian, navigation
HSGNSS	10 m	global	Parallel correlation, assisted GNSS	LBS
Pseudolites	cm–dm	10–1000	Carrier-phase ranging	Pit mines
Inertial navigation	1%	10–100	Dead reckoning	Pedestrian, navigation
Magnetics systems	mm–cm	1–20	Fingerprinting and ranging	Hospitals, mines
Infrastructure systems	cm–m	building	Fingerprinting, capacitance	Ambient-assisted living

Ubisense [29.15] and Time Domain [29.16] operate within restricted frequency bands (the FCC has restricted the frequency band for unlicensed UWB to 3.1–10.6 GHz while the European Communications Commission has restricted it to 6.0–8.5 GHz. UWB systems operate over short distances (typically ≈ 100 m depending on the operating environment and sensor configuration) and use TDoA, ToA, and signal travel time as measurement techniques to determine ranges between an UWB transmitter and receiver radios. UWB solutions are typically deployed as part of RTLS for efficient asset tracking, monitoring, and management. Other applications include relative positioning of vehicles and personnel tracking [29.17]. With positioning accuracies potentially at the levels of 0.2 m [29.18], UWB is becoming an important component of automotive, aerospace, and a variety of manufacturing processes.

Radio Frequency Identification Positioning. Similar to UWB, radio frequency identification (RFiD) positioning systems are primarily used to tag and track personnel and assets. They require the deployment of RFiD scanners across the operational environment. These scanners are then able to interrogate either active or passive tags attached to the object to be tracked. The range between the scanner and the tag is the most important relationship defining the positioning technique used (active tags enable a greater range than passive tags). COO and RSS ranging are the two most popular techniques used in RFiD positioning. For COO, the location of the reader is described by a cell identified by the maximum read range to a tag [29.19]. This technique offers relatively low accuracy and depends on the size of the distinguishable cells 10–20 m. A dense configuration of scanners across an area would improve the granularity of positioning but would incur a significantly higher cost. As a result, it is not practical to

use RFiD for real-time tracking applications over large areas. To improve the achievable positioning accuracy, the deduction of ranges to the RFiD tags from received signal power levels is used (it can be converted to a distance). Calibration for the signal strength to range conversion is required. The position fix can be obtained using trilateration if range measurements to several tags are performed. To create a more general localization method than using trilateration, the RFiD location fingerprinting is used. The principle of operation of RFiD fingerprinting is similar to that used in WiFi and cellular positioning systems. Significant efforts into improving the positioning accuracy of RFiD has demonstrated results at the meter level [29.20]. Commercial RFiD tags are now widely used in a range of safety and security critical applications.

Micro-Electro-Mechanical Systems (MEMS) Inertial Sensors. MEMS inertial sensors like gyros and accelerometers are a key enabling technology for LBS. These low cost, low profile motion sensors are currently embedded as standard in most modern mobile devices providing very reliable information about rotation and acceleration rates and – after integration – about relative velocity and position over a certain period of time (i. e., a few minutes depending on sensor type and quality) with a high frequency (> 100 Hz). These data ideally complement the long-term stable, low-rate GNSS measurements. Moreover, inertial sensors are suitable to bridge bad GNSS signal reception conditions for a couple of minutes and guarantee a constant accuracy of the hybrid positioning solution. MEMS inertial sensors have already demonstrated significantly improved performance from their first-generation configurations, with MEMS accelerometers approaching performances close to those of tactical grade inertial measurement units (IMUs) [29.21]. The short-term relative positioning accuracy of MEMS IMUs is very high.

Table 29.7 Performance specifications for different grades of inertial sensors

IMU sensor grade	Sensor name and components	Characteristics (1σ error coefficients)
Navigation grade	Honeywell HG9900 IMU Honeywell GG1320AN digital laser gyroscope	<ul style="list-style-type: none"> ● Bias: $< 0.003^\circ/\text{h}$ ● Random walk: $< 0.002^\circ/\text{Vh}$ ● Scale factor: < 5.0 ppm
	Honeywell QA2000 accelerometer	<ul style="list-style-type: none"> ● Bias: $< 25 \mu\text{g}$ ● Scale factor: < 100 ppm
Tactical grade	Systron Donner SDI500 MEMS quartz gyroscope	<ul style="list-style-type: none"> ● Bias: $1^\circ/\text{h}$ ● Random walk: $< 0.02^\circ/\text{Vh}$ ● Scale factor: < 200 ppm
	MEMS quartz accelerometer	<ul style="list-style-type: none"> ● Bias: $100 \mu\text{g}$ ● Scale factor: < 200 ppm
Consumer grade	Honeywell HG9000 IMU ADIS16334 iSensor MEMS accelerometer	<ul style="list-style-type: none"> ● Bias: $3^\circ/\text{s}$ ● Random walk: $2^\circ/\text{Vh}$
	ADIS16334 iSensor MEMS gyroscope	<ul style="list-style-type: none"> ● Bias: 12 mg

However, if unassisted, the errors and biases typical of these sensors accumulate rapidly, with positioning errors up to hundreds of meters [29.22]. Significant efforts in improving the performance of MEMS sensors through their integration with GNSS has resulted in much success commercially for land-based applications. In fact, for all of the alternative positioning technologies presented in this chapter, many successful commercial RTLS rely on a combination of multiple technologies to provide robust positioning capabilities in all environments. Table 29.7 lists the performance specifications for different grades of inertial sensors.

29.1.4 Intelligent Transport Systems

ITS can be defined as the application of advanced information and communication technology to surface transportation in order to achieve enhanced safety and mobility while reducing the environmental impact of transportation. Cooperative ITS (C-ITS) are an emerging capability that leverage wireless communication capabilities to enable vehicles and surrounding infrastructure to exchange information about the location, speed, and direction of other road users also using C-ITS [29.23]. Vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communications are supported by dedicated short range communication (DSRC) capabilities which enables two-way short-to-medium range wireless communications and permits very high data transmission – critical in communication-based active safety applications. The US FCC has allocated 75 MHz of spectrum in the 5.9 GHz band for use by C-ITS vehicle safety and mobility applications. DSRC offers significant opportunities for sharing of information between vehicles, pedestrians, and roadside infrastructure.

The positioning performance enhancements offered by multiconstellation GNSS, SBAS, and alternative positioning sensors, combined with the maturing of fundamental ITS technologies and the emergence of telematics driven C-ITS capabilities are facilitating applications well beyond those of traditional ITS including:

- Smart mobility applications improve the efficiency, effectiveness, and comfort of road transportation through [29.2]:
 - *Navigation*: The most widespread application, providing turn-by-turn indications to drivers through portable navigation devices and in-vehicle systems (IVS).
 - *Fleet management*: On-board units (OBUs) transmit GNSS positioning information through telematics to support transport operators in monitoring the performance of logistics activities.
 - Satellite road traffic monitoring services collect floating car location data from vehicles through PNDs, IVS, and mobile devices, processing this traffic information to be distributed to users and other interested parties.
- Safety-critical applications leverage precise and secure positioning in situations with potential harm to humans or damage to a system/environment:
 - While being connected vehicles GNSS positioning will be integrated with the information coming from other sensors and communication technologies in in-vehicle systems (IVS), enhancing the safety and comfort of the driver.
 - Dangerous goods tracking can be done by transmitting GNSS-based positioning data on the vehicles, carrying them along with other information about the status of the cargo.

Table 29.8 Positioning accuracy levels for C-ITS (after [29.24])

Type	Level	Accuracy requirement		Research prototype Root means square (order)	Communication latency (second)
		95% confidence level (m)	Root means square (order)		
V2I: absolute (V2I: Vehicle to infrastructure)	Road level	5.0	meter	Meter	1–5
	Lane level	1.1	Submeter	Submeter	1.0
	Where-in-lane-level	0.7	Decimeter	Decimeter	0.1
V2V: relative (V2V: Vehicle to vehicle)	Road level	5.0	Meter	Submeter	0.1
	Lane level	1.5	Submeter	Decimeter	0.1
	Where-in-lane-level	1.0	Decimeter	Centimeter	0.01–0.1

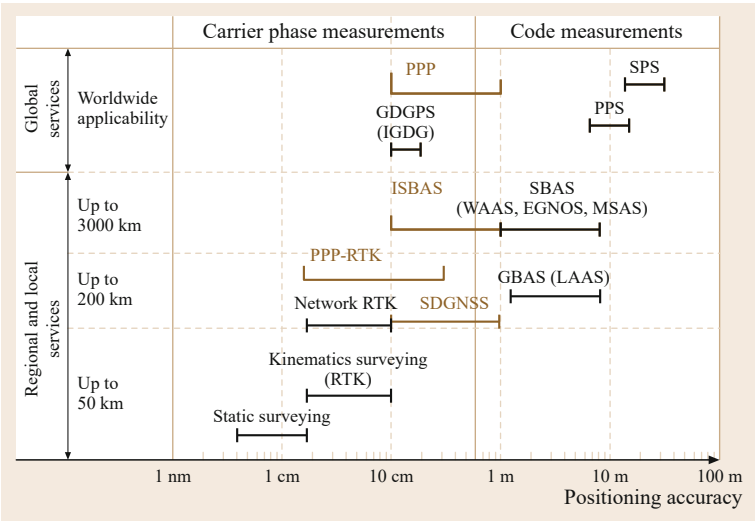


Fig. 29.9 GNSS performance levels (after [29.24]). Definitions for the acronyms contained in this figure are as follows: GPS, global positioning system; PPP, precise point positioning; SPS, standard positioning services; GDGPS, global differential GPS; IGDG, Internet-based global differential GPS; PPS, precise positioning services; ISBAS, integrated SBAS; SBAS, space-based augmentation system; WAAS, wide area augmentation system; EGNOS, European Geostationary Navigation Overlay Service; MSAS, multifunctional satellite augmentation system; RTK, real-time kinematic; GBAS, ground-based augmentation system; LAAS, local area augmentation system; SDGNSS, satellite differential GNSS.

- **Liability applications:** The positioning data provided by liability applications are linked to legal and economic liabilities:
 - In road user charging (RUC), GNSS-OBUs support toll operators in charging based on the actual use of the roads and in managing congestion control.
 - Insurance telematics black boxes rely on GNSS data to increase the fairness of motor insurance for both insurers and subscribers.
- **Regulated applications** apply the transport policies introduced by national or international legislations:
 - The GNSS-enabled IVS are used in regulated applications, such as the European eCall or the ERA-GLONASS in Russia, which send an emergency call to 112 in the case of an accident, thus accelerating emergency assistance to drivers.
 - Enhanced digital tachographs leverage GNSS positioning to support road enforcers, recording the position of a given vehicle at different points during the working day.

In the following section, the positioning requirements of C-ITS applications are presented. Following this, the positioning techniques used to achieve these performance levels are also discussed followed by a discussion on the benefits of DSRC for C-ITS and how the position solution can be improved by using cooperative or collaborative positioning techniques.

GNSS Performance Classification of C-ITS Applications

C-ITS applications cut across a range of GNSS performance parameters, with low-end applications such as vehicle tracking typically requiring low cost and low accuracy positioning information while high end applications such as lane control require high accuracy and high availability information. Increasingly, integrity for ITS applications is becoming more relevant as safety-critical applications gain prominence, for example, collision avoidance and autonomous vehicle navigation. The positioning accuracy requirements for C-ITS safety applications are typically classified into three levels: road-level (on which road the vehicle is placed); lane level (in which lane the vehicle

Table 29.9 Summary of GNSS-based positioning techniques in terms of accuracy and their related C-ITS applications (after [29.24]). Definitions for the acronyms contained in this table are as follows: GPS, global positioning system; GNSS, global navigation satellite system; PPP, precise point positioning; PPS, precise positioning service; SBAS, space-based augmentation system; RTK, real-time kinematic; WADGPS, wide area differential GPS; DGPS, differential GPS

Tier	Technique option	Status		Accuracy range	Cost	C-ITS applications
		Current	Future			
1	A	Standalone GPS (SPS)	Standalone multiple GNSS (use of Multi-GNSS in standalone mode is inherently more reliable than GPS alone)	10–20 m	Low	Vehicle navigation, personal route guidance and LBS
2	A	Standalone GNSS (PPS). Code DGPS	Standalone multiple GNSS positioning	1–10 m	Low	Vehicle navigation, LBS, road traffic management
3	B	Current WAAS Commercial WADGPS	Future SBAS design for multiple-GNSS	0.1–1 m (utilizing SBAS and V2V relative positioning)	Low	C-ITS safety applications: lane-level positioning, lane-level traffic management, and where-in-lane-level applications
4	C	Smoothed DGPS	Smoothed DGNSS	0.1–1 m	Medium	Research prototype C-ITS safety systems, offering bench mark solutions for testing low-cost units
	D	RTK	Combined PPP and RTK (seamless)	0.01–0.1 m	Medium to high	
	E	PPP				
5	Advanced D and E	Static positioning	Subcentimeter RTK with multi-GNSS Signals	0.001–0.01 m	High	Geosciences and geodynamic studies. Not recommended for C-ITS applications

is in); where-in-lane-level (where the vehicle is in the lane). In Table 29.8, these levels have been quantified. From this, a classification of positioning accuracy levels for C-ITS can be represented as: 10–20 m; 1–10 m; 0.1–1 m; 0.01–0.1 m and 0.001–0.01 m. It is useful to note that the full positioning performance requirements for emerging C-ITS applications are yet to be specified and in many cases are still being fully defined.

Figure 29.9 presents the accuracy levels of existing and established GNSS positioning services. From this, the GNSS-based techniques that could be used for positioning in C-ITS applications are as follows:

- **Technique A:** Standalone (global navigation satellite system) GNSS absolute positioning and V2V relative positioning with low cost GNSS receivers.
- **Technique B:** Space-based augmentation system (SBAS) absolute positioning and/or V2V relative positioning with low cost GNSS receivers.
- **Technique C:** Smoothed differential GNSS (DGNS) absolute positioning and/or V2V relative positioning with low cost GNSS receivers.

- **Technique D:** Real-time kinematic (RTK) positioning with dual-frequency receivers.
- **Technique E:** PPP and V2V relative positioning with high-end GNSS receivers.

Table 29.9 presents a summary of the five-tier GNSS positioning techniques and their relation to C-ITS applications.

Cooperative Positioning and Sensor Fusion for C-ITS

As road transport applications become increasingly safety-liability-critical, the requirements for better positioning accuracy, availability, and integrity have motivated the trend toward the development of positioning systems that combine measurements from a range of traditional and nontraditional signals. Systems such as mobile phone devices simply switch between available alternatives when GNSS is unavailable. More robust techniques aim to determine the optimal solution based around the use of estimation techniques. Systems that combine GNSS measurements with low-cost inertial

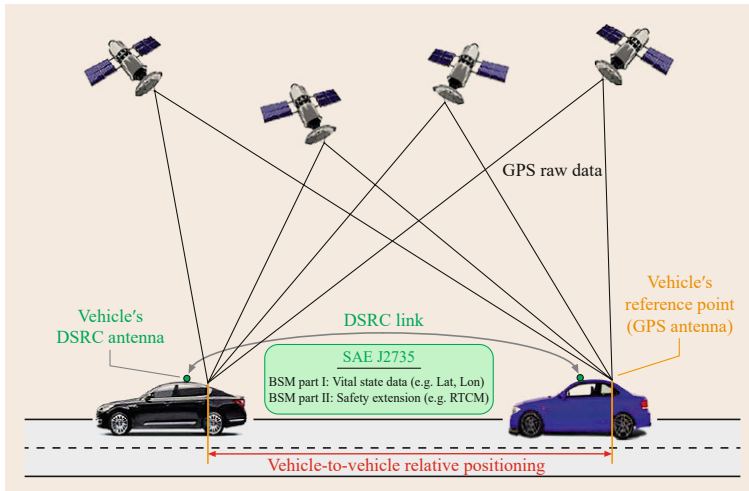


Fig. 29.10 DSRC-enabled CP (after [29.24])

sensors are available with different levels of computational complexity. While many of these are aimed at high-end applications, the cost and size reductions in these sensors are seeing them embedded in mobile devices and consequently now routinely integrated in positioning and navigation applications.

Collaborative or cooperative positioning (CP) techniques have been adopted from the field of wireless sensor networks as an approach to improving the navigation and positioning performance for a range of human and land vehicle navigation applications. This is particularly relevant for those applications operating in GNSS-challenged environments where requirements for positioning availability cannot be met and/or which are safety-critical, requiring higher levels of reliability and integrity. CP techniques typically leverage an available communication infrastructure to share information between users operating within a defined neighborhood or so-called ad hoc network. This shared information can be integrated to deliver more robust positioning performance. Under certain conditions, the communication infrastructure itself can be used as a measurement source for positioning. Figure 29.10 shows the concept of relative positioning with GNSS raw data exchanged between vehicles using DSRC communications capabilities.

DSRC has the potential to provide a ranging measurement between vehicles in a vehicular ad hoc net-

work (VANET). What is emerging as a significant consideration for CP are the benefits for positioning in terms of availability, integrity, reliability, and accuracy versus cost in terms of infrastructure, computational overheads, and the overall quantity and quality of information that needs to be shared to meet the positioning requirements of a specific application. CP algorithms typically use estimation and in particular decentralized filtering and tracking models to combine low-cost positioning sensors and signals, map matching, and DSRC.

Fusion of measurements from a number of sources can be approached in a Bayesian estimation framework. Here, the aim is to compute the distribution of an object state, which includes the object position and velocity, conditional on the measurements received from all sources. This is referred to as the posterior distribution. Because the relationship between the measurements and the object state is nonlinear, the posterior distribution cannot be found in the closed form. Many approximations have been proposed to address this problem. The most popular approximation is to linearize the measurement equation about the a priori expected value of the object state. This leads to the well-known extended Kalman filter (EKF; Chap. 22). This approximation is not always reliable but has been shown to provide good performance in many PN applications.

29.2 Rail Applications

The rail sector is recognized as a key contributor to realizing more efficient, safer, and sustainable mobility and ITS in the future. It is also widely accepted that knowledge about the position of each rail vehi-

cle in the route network is compulsory for both safety and nonsafety relevant applications, and that the position and timing information provided by GNSS, when combined with other sensor systems, communication,

and information technology infrastructures, are critical to reducing accidents, delays, and operating costs while increasing track capacity, customer satisfaction and cost effectiveness [29.25]. However, despite the potential benefits offered by GNSS, liability, safety, and operational concerns surrounding current GNSS performance have resulted in a significant lag in railway ITS developments when compared to the land and maritime sectors [29.26].

Applications across the railway sector can be broadly classified according to the positioning performance required to support the individual application requirements, with safety-critical applications typically requiring higher levels of positioning availability, integrity and accuracy than non safety-critical applications. In [29.27], to investigate the performance and quality levels of GNSS positioning as required for railway applications, the European Rail Advisory Forum has proposed three main classes of applications, safety, operational, and professional. A summary of

these applications and their positioning specifications is presented in Table 29.10. Within this classification, a description of relevant railway applications is presented in the following sections.

29.2.1 Signaling and Train Control

The principal motivation for train control is to prevent collisions when trains are traveling on the same track – either in the same direction (following each other) or traveling in opposite directions (toward each other). It is the safety critical part of train control, in which, automatic train control (ATC) systems continuously monitor all train movements and provide fail-safe signaling. For example, the ATC keeps tabs on a train's speed as it heads toward curves, automatically adjusting it if the driver fails to do so. The next generation of ATC technology is now referred to as communications-based train control (CBTC). A CBTC system is a [29.28]:

Table 29.10 GNSS positioning requirements for different classes of railway applications (TBD: to be defined; ELM: European Land Mass)

Application	Requirement							
	Horizontal accuracy ^a (m)	Integrity alert limit ^b (m)	Integrity max. time to alarm ^c (s)	Availability ^d (%)	Service interrupt (s)	Continuity ^e (%)	Coverage ^f	Fix rate (s)
Safety-related applications								
ATC on high density lines	1	2.5	< 1.0	> 99.98	< 5	> 99.98	ELM	1
Train control on medium density lines	10	20	< 1.0	> 99.98	< 5	> 99.98	ELM	1
Train control on low density lines	25	50	< 1.0	> 99.98	< 5	> 99.98	ELM	TBD
Mass commercial/information and management – operational applications								
Tracking and tracking of vehicles	50	125	< 10	99.9	TBD	TBD	ELM	TBD
Cargo monitoring	100	250	< 30	99.5	TBD	TBD	ELM	TBD
Dispatching	50	125	< 5	99.9	TBD	TBD	ELM	TBD
Passenger information	100	250	< 30	99.5	TBD	TBD	ELM	TBD
Infrastructure and civil engineering, professional application								
Positioning of machines	0.01	TBD	< 5	99.5	TBD	TBD	Operating area	TBD
Infrastructure survey	0.01	10 ⁻³	< 10	99	TBD	TBD	ELM	TBD
Fix point applications	0.005	TBD	< 30	99	TBD	TBD	ELM	TBD

^a Accuracy is specified as the position error at 95% confidence level.

^b Threshold value or alert limit – the maximum allowable error in the measured position before an alarm is triggered.

^c Time-to-alarm – the maximum allowable time between an alarm condition occurring and the alarm being present at the output.

^d Defined as intrinsic availability: This is the *Probability that a system or equipment is operating satisfactorily at any point in time when used under stated conditions, where the time considered is operating time and active repair time.*

^e Continuity is defined as the probability that the location unit will be able to determine its position within the specified accuracy and is able to monitor the integrity of the determined position over the mission time, in all points of the route within the coverage area.

^f The coverage is defined as the surface area or volume of space where the SIS service is sufficient to permit the user to determine its position with the specified accuracy and to monitor integrity of the determined position.

continuous, automatic train control system utilizing high-resolution train location determination, independent of track circuits; continuous, high-capacity, bidirectional train-to-wayside data communications; trainborne and wayside processors capable of implementing automatic train protection (ATP) functions, as well as optional automatic train operation (ATO) and automatic train supervision (ATS) functions.

CBTC includes three major components:

- Automatic train protection (ATP) assures that safety of trains by preventing collisions and derailments. It ensures that trains are separated by a safe distance, prevents overspeeding, and conflicting movements at junctions and crosses.
- Automatic train operation (ATO) controls the movement of the train specifically regulating speed, station stoppings, and opening and closing of doors.
- Automatic train supervision (ATS) assigns routes, monitors running trains, and provides data so that their service may be adjusted by controllers to minimize delays.

The primary functions of these subsystems can only be delivered through timely (real-time or near real-time), accurate, and continuous knowledge of where the train is on the track. The conventional technique for determining train location is through *block occupancy*. The conventional principle is that the track is divided into fixed *blocks* (a section of length equal to the distance needed for the train to come to a full stop under ordinary braking conditions and with a certain safety margin [29.29]). Only one train is allowed into a block at a time (Fig. 29.11).

Track circuits and axle counters are still widely used for ATP, to detect the presence of a train in a particular block. A track circuit is a simple electrical device used to detect the absence of a train on rail tracks. *Axle counters* refer to subsystems that are utilized for track vacancy detection on fixed-guide way transport systems such as railways. While the axle counter counts the vehicle axles entering and leaving a track section by

means of the so-called rail contacts and indicates the track section as occupied when the number of axles counted at the exit differs from the number of axles counted at the entrance [29.30]. All ATC and ATP functions are based on this very coarse position determination.

To provide finer positioning resolution, the *moving block* system underpinning CBTC does not require traditional fixed-block track circuits for determining train position. Instead, it adapts to the current situation, taking into account its own accurate speed and position as well as that of any train ahead or behind. In this way, the minimum distance between two trains following each other is reduced to the braking distance of the second. Safety and reliability is therefore highly dependent on the accuracy and integrity of the position and speed measurements.

Railways provide a difficult operational environment for GNSS technologies. Tunnels, cuttings, and urban canyons are all problematic factors, potentially causing blockage of GNSS signals thus degrading GNSS availability and integrity. GNSS is currently unable to meet the specified integrity threshold for positioning for train control applications of 10^{-9} failures per hour. This is due to a failure risk of 10^{-4} /h for any satellite under SPS [29.32]. Both of these contribute to the low uptake of GNSS for safety-critical rail systems. Figure 29.12 shows the applications requirements compared to the integrity levels available from GPS/Galileo/EGNOS.

To address the limitations of GNSS for rail applications, it is acknowledged that it must be hybridized with other sensors to determine a position sufficiently accurate for an use in safety applications. The Grail-2 (2007–2013) project funded by the GSA had the primary objective to develop a GNSS-based odometer, which would improve the ability of a train to determine its own position, integrating, and/or replacing the traditional odometric systems currently used in the ETCS (European Train Control System, a component of the European Railway Traffic Management System (ERTMS)) environment (tachometers, inertial navigation system (INS), Doppler radar, etc.) termed

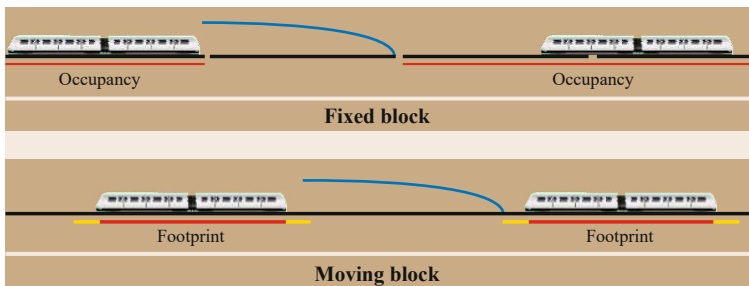


Fig. 29.11 Block signaling for ATP. Image courtesy of Israel.abad/Wikimedia Commons distributed under the CC BY-SA 3.0 license

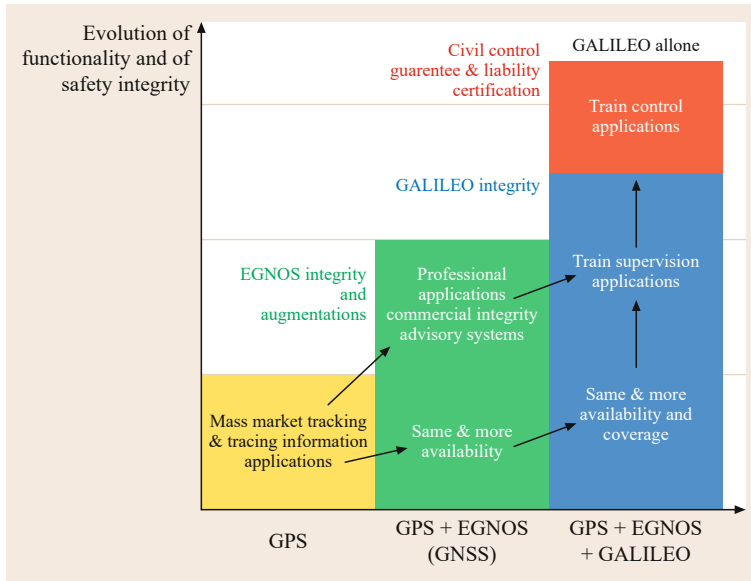


Fig. 29.12 Satellite navigation systems and railway applications (after [29.31])

enhanced odometry, it integrated GNSS as an additional sensor to compensate for odometry problems in high-speed runs (slip and slide phenomena) [29.33]. Standard odometry approaches to estimating the train position within a block has relied on an on-board odometry subsystem that computes the distance traveled and therefore the updated position of the train – the so-called dead reckoning approach. Any errors in the distances due to wheel slip and slide can only be corrected with a reset from the next transponder that the train passes [29.34].

Under the GRAIL and GRAIL-2 projects, a dual frequency GNSS receiver augmented by EGNOS, was

integrated with an IMU. The filtering and fusion of data coming from different sensors are performed by a Kalman filter – so-called data fusion filter (DFF), the GRAIL-2 implementation of the Kalman filter, is shown in Fig. 29.13, where

- PVA receives the GNSS position and velocity and the IMU measurements.
- DSA exploits the speed measure provided by the GNSS and the along track measured given by the IMU.
- RW receives measurements of angular rate by IMU and direction from two well-spaced GNSS antenna.

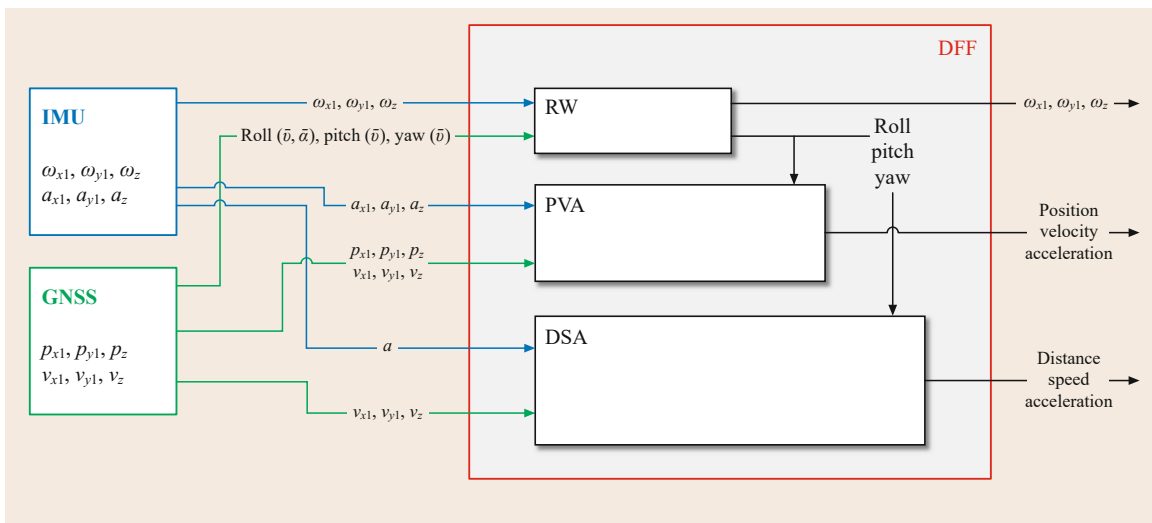


Fig. 29.13 GRAIL-2 measurement fusion for enhanced odometry (after [29.33])

The InteGRail (for INTElligent inteGRation of RAILway) (2005–2008) was another European project aimed at providing train position, velocity, and heading to the following specification: accuracy: 10 m along-track and 1 m cross-track (2σ) for discrimination of parallel tracks. The InteGRail prototype integrated measurements from an L1 GNSS/EGNOS receiver with an odometer, along-track acceleration sensor, fiber optical sensor for measuring the vertical rotation axis (i.e., the azimuth or heading angle), and digital route map [29.35]. More recently, the 3InSat (Train Integrated Safety Satellite System) (2012–2015) project funded by STS Ansaldo and the European Space Agency (ESA) aimed to develop, test and validate in a real set up a new satellite-based platform suitable for a train control and management system [29.36].

However, GNSS when combined with other systems, such as IMUs, does have the potential to provide high accuracy and integrity positioning for train control/rail management systems. Positive train control (PTC) initiatives in North America and many other countries, as well as steps toward a European Railway Traffic Management System (ERTMS) will result in significant growth in the uptake of GNSS across this sector. For positioning, the 3InSat project is aimed at designing and developing a multisensor location detection system (LDS) using multiconstellation GNSS.

The LDS multisensors/multiconstellation (GPS, GLONASS, Galileo, BeiDou) system which, through

association of different SatNav receivers and on-board sensors (gyros, accelerometers, tachometer) will provide a safety compliant positioning solution able to reach high integrity values and improve the overall availability and resiliency. The LDS will make use of SBAS (e.g., EGNOS for Europe) in combination with GBAS for both differential corrections and integrity monitoring. Additionally, the LDS will have independent integrity monitoring on-board capability to further mitigate GNSS errors and autonomously assess the GNSS location integrity in the case of augmentation data unavailability (e.g., EGNOS SIS unavailability). A track area augmentation and integrity monitoring network is to be installed along the railways tracks. This element will be mainly used in regions out of the SBAS footprints.

Figure 29.14 shows the reference architecture for the 3InSat systems. The primary task of the space segment is to provide the reference satellite signals needed for train position computation as well as to distribute real-time corrections related to satellite ephemerides, clock offsets, propagation delays, and signal-in-space (SIS) integrity. The (track area) augmentation and integrity monitoring network plays a role similar to the EGNOS range and integrity monitoring subsystem and, in fact, it will be deployed only on those areas out of EGNOS footprint. The on-board unit through the localization determination system (LDS) subsystem computes the train position by using the GNSS signal,

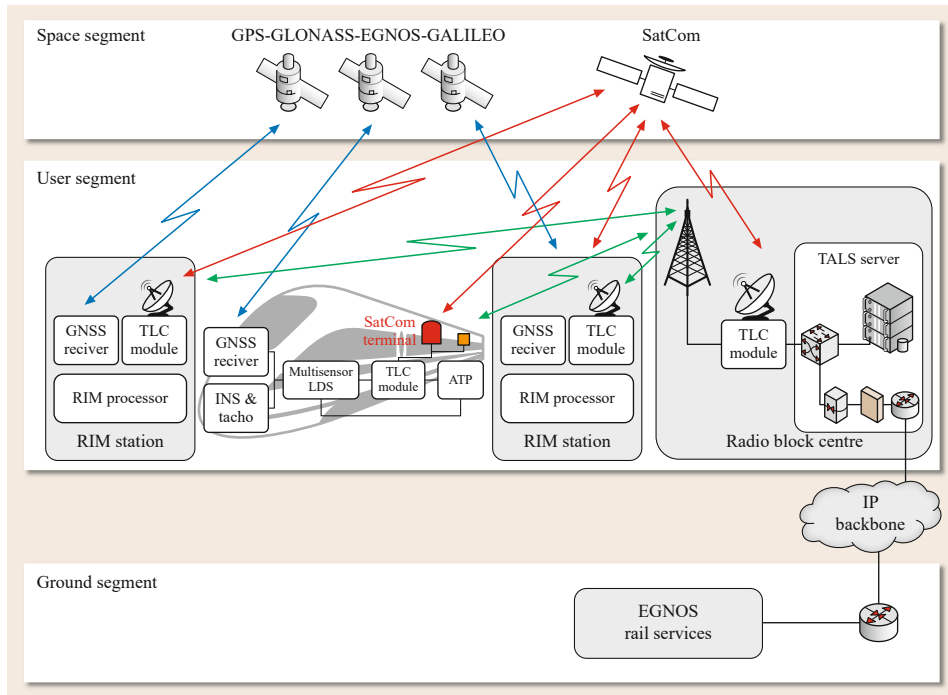


Fig. 29.14
3inSAT reference
architecture
(after [29.37])

the augmentation information for integrity monitoring, and the data from other sensors as inertial navigation systems (INS) and tachometers. The on-board bearer-independent telecommunication subsystem will take care of train control messages and other important system information [29.36].

Infrastructure Data Collection

Train protection delivered through PTC in North America and the proposed European Train Control System (ETCS) monitors the train speed against the current permitted speed limit. The speed may be limited by line profile or signal indication, that is, the need to protect routes of other trains and track-related constraints. If the allowable speed is exceeded, a brake application is invoked until the speed is brought within the required limit or the train is stopped. Each block of the track is described by a fixed dataset related to its location, length, gradient and maximum speed limit, works on the line, etc., and communicates this to the train as it passes. Of many challenges that face railroads to implement this complex system is the need for a highly accurate base map and asset inventory database which must be compatible with all other PTC/ETCS systems. A solid base-map is the foundation upon which a PTC or ETCS implementation is built upon in order to properly control train movement based upon track slope, curvature, and speed zones.

Many commercial solutions exist for generation of the required spatial database for PTC/ETCS. Perhaps the most popular is the mobile mapping solution based around mobile Lidar, dual frequency, high precision GNSS, and on-board INS. These systems overcome limitations in GNSS-denied environments through the use of a hybrid positioning capability, offering better than 5 cm positioning accuracy and a resolution of up to 1 cm [29.38]. For example, in support of PTC/ETCS, the commercial Sanborn mobile mapping system can provide:

- Accurate knowledge of train location and GNSS for signal and braking control.
- Create a complete inventory of assets, obstructions, clearances, switches, frogs, crossings, etc.
- The complete environment of the crossing, including the condition of railway cross guards (if any), allowing for the maintenance of these guards; this data has potentially critical value to railroad, insurance, and forensics agencies if an accident occurs.

Every year, railway custodians globally spend millions of dollars to inspect the rails for internal and external flaws. Testing/inspection systems need to be synchronized with a positioning system, and here GNSS-based

mobile mapping systems can provide more accurate data than traditional techniques. Secondly, recent developments in remote monitoring system (which are typically based on a combination of GNSS/GIS (geographic information system) technologies, wireless communications, signal processing, and embedded computing) have become compact and relatively inexpensive, and can therefore be used on almost any rail vehicle. These kinds of solutions that enable real-time performance monitoring contribute significantly to safety since they allow quicker detection of defects than traditional methods which are based on separate periodical track inspections.

The Alstom TrainTracer is one example of a commercially available module that can be retrofitted to the train's control system, which remotely monitors changes in its major components and reports wirelessly to a ground-based server. By fitting the condition monitoring system with GNSS location tracking to each train, fault data can be processed and transmitted while the train is in service, rather than at scheduled maintenance times. Engineers can therefore remotely analyze and diagnose the on-board scenario facilitating more efficient detection of problems or troubleshooting failures. In addition, the Sanborn Automatic Track Inspection Program (ATRIP) provides the following key features:

- High relative accuracy between rails of 7 mm
- Meets positioning accuracy clearances
- Point cloud density is great enough to obtain all relevant information in one pass minimizing needed track time
- Mobile mapping data can be integrated with other data (ortho, LiDAR, etc.) to build a complete picture of rail right-of-way.

Level Crossing Protection

All types of train protection systems are based on the desire to reduce or eliminate the possibility of driver error resulting in a train-movement-related accident by failing to obey a visually displayed line-side or in-cab signal instruction. Level crossing protection systems are actuated by knowing the position and speed of the train on the track. In a conventional automatic level crossing (ALX), the control system includes fixed sensors along the track which can detect a train approaching the crossing [29.39]. It senses not only the approach of the train but also the speed so that if a train is coming very slowly they will not activate until the train is close. When the traffic signals are activated, shortly after, the barriers are closed. Once the road traffic is stopped, the railway signal controlling the route over the crossing can be set to allow the train to proceed. The distances of the sen-

sensor and signal from the crossing are determined by the highest speed at which trains are permitted to travel on the section of line in question. The distance from the sensor to the signal must be great enough for the ALX to fully operate before the train reaches the signal. If the ALX is not ready for the train to cross, the train must be able to stop in the distance between the signal and the crossing. Slower trains cause the crossing barriers to close unnecessarily early, increasing the waiting time for road traffic.

As a safety application for railway, much of the hybrid positioning capabilities proposed for ATC functions can be leveraged into triggering a warning on the approach to a level crossing, with a consistent time lapse regardless of the speed of the train. The EGNOS Controlled RAILway equipment (ECORAIL) project is a demonstration of the capabilities of GNSS for ALX protection. This project developed an on-board unit that is able to determine the position of a train on the track with sufficient accuracy and reliability to be used for railway control purposes. A GNSS receiver augmented with EGNOS, odometer measurements, and a track database was used to demonstrate the control function of closing the barriers at the level crossing as a train approaches.

For the first part of the demonstration, the performance of ECORAIL was assessed as the on-board system computed the location of the train and determined the position on the track at which the crossing barriers should be closed. The closure commands were transmitted by radio link to a ground station, where their time-of-arrival was recorded. The ground equipment also recorded the times at which all the events in the conventional control equipment occurred. The positional accuracy achieved by ECORAIL was better than 3 m when compared with the fixed sensors. For the second phase of the demonstration, ECORAIL determined the timing of the barrier closure commands based on both the position and speed of the train, allowing the ALX operation to be optimized – slower moving trains are able to close the crossing when they are nearer to it, rather than relying on the fixed detectors.

Subsequent projects in a similar direction have identified significant performance benefits from a hybrid onboard system for position and speed determination. In almost all cases, these solutions represent the state of the art in positioning in difficult railway environments for safety critical applications.

29.2.2 Freight and Fleet Management

The ability to effectively track the location of goods and to estimate their delivery times is as important in the railway sector as in other modes of transporta-

tion. Similarly, accurate and timely information as to where a train is or has been is paramount to ensure efficient operation and service passenger information systems. A significant difference for the rail industry compared to other vehicular fleet management system is that [29.40]:

the state of the art in tracking and managing key rolling assets – railcars, tankers, and locomotives – lags far behind other highly capitalized, mission-critical industries.

As with other positioning and tracking applications, the positioning capabilities of DGNSS – operating under ideal operating conditions can deliver the positioning performance required for robust tracking and monitoring functions. It is in difficult environments that positioning performance and consequently application service levels are compromised. As with many other similar applications, the use of hybrid solutions are an effective approach to providing a ubiquitous positioning capability. The hybrid positioning technologies used to position the train on the track is sufficient to meet the requirements for many freight and fleet management functions. Other technologies that are being deployed to help achieve the full benefits of a railway RTLS include: RFiD tags [29.41] and UWB. Ubisense has introduced an UWB RTLS that automates yard operations and improves efficiency and productivity. By precisely locating rolling stock in both indoor and outdoor areas, operators gain an unprecedented level of visibility into workflow which helps us to improve all areas of rail yard operations.

Benefits as documented by a commercial asset management company, RFTrax, include:

- Owners of large railcar fleets could use these data to significantly reduce the time spent in stations by being able to monitor the exact location of an entire railcar fleet, thus preventing *warehousing* of railcars and speeding up their reuse.
- Shippers could use these data to improve overall service delivery and efficiency by better managing railcar movements and use.
- The railroads themselves could significantly lower their costs and improve their margins by being able to understand the sources of problems such as truck hunting – the potentially destructive shifting of railcars and their loads in transit – as well as derailments and improve their maintenance of both their track and rolling assets.
- Integrating those data *upstairs* into the enterprise resource planning (ERP) systems can provide an even greater return on investment (ROI): Integrating rail-

car data with ERP data can significantly improve the rail freight industry's ability to be a key partner in initiatives such as just-in-time or lean manufacturing, vendor managed inventory, and other highly visibly efforts to improve supply chain responsiveness across the board.

29.2.3 Passenger Information Systems

An inevitable outcome of high accuracy Real Time Kinematic (RTK) positioning and monitoring of vehicle fleets has been the ability to provide current updates on arrival and departure schedules to passengers. Over the last decade, passenger information systems (PIS) have evolved from simple standalone audio and visual displays to multimodal integrated systems that keep passengers informed, safe, and entertained along their journey in public transit systems (metro trains, commuter rail, station platforms, buses, or bus shelters) [29.42]. PIS typically deliver real-time information seamlessly on-board vehicles, through web browsers and mobile devices, or physically at a station or transportation hub, while controlled and managed from a single control center. Many of the services provided by PIS, either pre-trip or on-trip, require the real-time position of the train to be monitored and reported continuously. Typical pre-trip information provided by a PIS includes the

service provided by the next vehicle to arrive, the expected and timetabled arrival time of the next vehicle. On-trip information includes the name of the next station, the time of arrival there and advice on connecting services.

PIS tend to be multimodal with the most popular implementations connecting the primary forms of land transport (buses, trains, and trams) to ensure seamless travel across heavily trafficked routes. In all cases, an onboard system is required to determine and transmit the geographical location of the train. For buses, GNSS is primarily used for this task. In the rail industry, a number of leading electronics manufacturers are now offering PIS technology that can be retrofitted to existing railway vehicles. These systems either interface with GNSS directly or access positioning information from the train management system. PIS typically have low positioning accuracy requirements but in order to provide an enhanced user experience, this information needs to be available and reliable. The positioning information provided needs to be an order of magnitude higher than the underlying service. *Thus, if 99% of the trains run on time, the service information presented to users needs to be 99.9% accurate* [29.43]. In many cases where GNSS cannot satisfy this requirement, augmentation sensors and fusion techniques are necessary.

29.3 Maritime Applications

The maritime sector was an early adopter of GNSS for general navigation applications, with GNSS now a mandatory navigation aid for all ships regulated under the International Maritime Organization (IMO), International Convention for the Safety of Life at Sea (SOLAS) maritime safety treaty. The navigation requirements of non-SOLAS regulated vessels such as merchant, commercial, and recreational vessels, also depend heavily on GNSS. In addition, GNSS is used to ensure safe navigation in inland waterways (IWW) such as rivers, canals, lakes, and estuaries. The enhanced navigation performance from future multi-GNSS constellations is expected to deliver GNSS receivers that meet IMO performance standards for maritime operations. Specifically, GNSS capabilities are expected to underpin the IMO e-navigation initiative which aims to increase the safety and security of maritime navigation through the integration of all navigation tools into a comprehensive system that can collect, integrate, exchange, present, and analyze marine information on ships and at shore. Resilient position, navigation, and timing (PNT) has been identified as a core requirement for the successful implementation of the e-navigation concept [29.44].

IMO Resolution A.915(22) differentiates between the performance requirements of GNSS for navigation and positioning applications. Maritime positioning applications include vessel monitoring, traffic management, port operations, search and rescue (SAR), marine engineering, etc., and are representative of the increasing range of GNSS applications aimed at improving the safety and productivity of maritime operations. Figure 29.15 shows the expected growth (to 2023) in GNSS receiver shipments based on applications in the maritime sector. Much of this trend is being driven by the increased uptake in GNSS technology by the maritime consumer/recreation market, improved operational efficiency, portability, and durability of SAR technologies and policies that mandate capabilities for vessel monitoring systems (VMS) and automatic identification systems (AIS).

The increasing significance of GNSS to the maritime sector combined with concerns surrounding the integrity and vulnerability of positioning for safety and liability critical applications has established a vision for resilient maritime PNT. It is envisaged that GNSS or augmented GNSS will be the core PNT capability

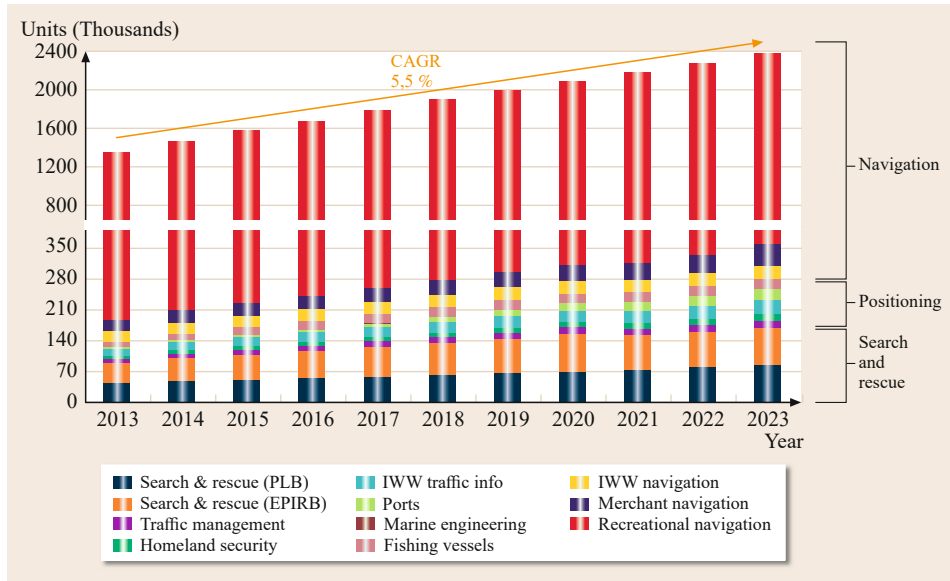


Fig. 29.15 Maritime applications projected growth (2013–2023) (after [29.2], courtesy of European GNSS Agency)

underpinning maritime navigation and positioning applications in parallel with independent, dissimilar, and complementary positioning systems depending on the needs of the specific application.

In this section, GNSS performance capabilities for the maritime sector are presented. Augmentation systems that improve the fundamental performance are also discussed within the context of high-performance maritime applications including navigation systems, fishing vessel control, SAR, and port operations.

29.3.1 GNSS Performance Requirements for Maritime Applications

The requirements of maritime positioning and navigation applications are very diverse, but fundamentally rely on the output from an electronic position fixing system (EPFS), in order to effectively perform their required tasks. GNSS has become the primary means of providing EPFS information with GPS, GLONASS, and BeiDou endorsed as meeting IMO requirements as a World-Wide Radio Navigation System (WWRNS). IMO Resolution A.915(22) provides a benchmarking set of operational performance requirements (in terms of accuracy, coverage, availability, continuity, and integrity measures) for future GNSS to be endorsed as a WWRNS [29.45]. Table 29.11 shows the IMO GNSS performance requirements for general navigation.

In addition, IMO Resolution A.915(22) lists more than 30 different marine positioning applications and their proposed requirements – with accuracies ranging

from 10 cm to 10 m. In [29.46] to facilitate a comprehensive analysis of the performance capabilities of future GNSS from this list, applications were grouped into sets defined through similar recommended navigation performance (RNP) parameters (Tables 29.12 and 29.13). This results in 12 groups of applications. The differentiators between groups are: whether two-dimensional or three-dimensional position fixes are needed; the accuracy needed and the associated integrity alert limit the coverage, in terms of global, regional (continental), or local whether or not continuity is specified as a requirement. All groups of applications have common requirements for availability (99.8%), integrity risk ($10^{-5}/h$) time-to-alarm (10 s), and update rate (1 Hz). To achieve these levels of positioning performance, for the most safety critical of applications, robust PNT information requires three complementary components: a core GNSS; augmentation of GNSS to ensure that GNSS performance is fit for purpose; and adequate backup in the event of GNSS system failure.

In [29.46] the NEMO (Navigation system analysis for European Maritime Operations) software suite was designed to analyze the performance of different GNSS constellations and combinations of different GNSS augmentations. The tool was designed with the specific objective to analyze performance using parameters, metrics, and definitions applied in the maritime sector. Table 29.14 summarizes the predicted capabilities of GNSS scenarios to meet the requirements of the 12 application groups.

Table 29.11 IMO GNSS performance requirements for general navigation based on resolution A.915(22) (after [29.47])

	System level parameters				Service level parameters			
	Absolute accuracy	Integrity			Availability (% per 30 days)	Continuity (% over 3 h)	Coverage	Fix interval (s)
	Horizontal (m)	Alert limit (m)	Time to alarm ^b (s)	Integrity risk (per 3 h)				
Ocean	10 (100) ^a	25	10	10 ⁻⁵	99.8	N/A	Global	1
Coastal	10	25	10	10 ⁻⁵	99.8 (99.5)	N/A (99.85)	Global	1
Port approach and restricted waters	10	25	10	10 ⁻⁵	99.8 (99.8)	99.97 (99.97)	Regional	1
Port	1	2.5	10	10 ⁻⁵	99.8	99.97	Local	1
Inland waterways	10	25	10	10 ⁻⁵	99.8	99.97	Regional	1

^a Figures in brackets refer to operational requirements according to Res. 953^b More stringent requirements may be necessary for ships operating above 30 knots**Table 29.12** Positioning performance requirements of future GNSS for maritime transport applications (2-D) (after [29.46])

Accuracy/alert limit	Continuity	
	Not specified	Specified (99.97% over 3 h)
10 m/25 m	<i>Group 1</i>	<i>Group 4</i>
	Global	Global
	Ocean navigation	Track control
	Coastal navigation	Ship-to-ship coordination
	Search and rescue	Regional
	Casualty analysis	Port approach
	● Ocean	Inland waterways
	● Coastal	Ship-to-shore coordination
	Fisheries	Shore-to-ship management (coastal VTS)
	● Location of fishing grounds	
	● Positioning during fishing	
	● Yield analysis	
	● Fisheries monitoring	
1 m/2.5 m	<i>Group 2</i>	<i>Group 5</i>
	Regional	Local
	Cable and pipe laying	Port navigation
	A to N management (current)	Tugs and pushers
	Casualty analysis	Icebreakers
	● Port approach	Local (port or inland waterway) VTS
	Offshore exploration and exploitation	
	● Exploration	
	● Appraisal drilling	
	● Field development	
	● Support to production	
	● Postproduction	
0.10/0.25 m	<i>Group 3</i>	<i>Group 6</i>
	None	None

Table 29.13 Positioning performance requirements of future GNSS for maritime transport applications (3-D) (H) – horizontal position, (V) – vertical position, (S) – speed (after [29.46])

Accuracy/ Alert limit	Continuity	
	Not specified	Specified (99.97% over 3 h)
10 m/25 m	Group 7	Group 10
	Global	Global
	Oceanography	None
1 m/2.5 m	Group 8	Group 11
	Local	Local
	Dredging	Automatic docking
	Construction works	
	Local	Local
0.10 m/0.25 m	Container/cargo management	Port navigation (future)
	Law enforcement	
	Group 9	Group 12
	Local	Local
	Dredging	Automatic docking
	Construction works	
	Cargo handling	

Table 29.14 Summary of the predicted capabilities of GNSS scenarios

Appli- cation group	Degree to which scenario meets requirements					
	GPS (present)	GPS and IALA DGPS (present)	GPS and EGNOS (near future)	GPS and Galileo (future)	GPS and multiple freq. DGPS	GPS and carrier-phase DGPS
1	✓	In coastal areas within beacon coverage	Within the European Maritime Area	Globally	In coastal areas within beacon coverage	In coastal areas within beacon coverage
2	✓	Within very restricted area	✓	✓	In coastal areas within a range of 100 km from reference station	In coastal areas within beacon coverage
3 (null group at present)	✓	✓	✓	✓	✓	Within close proximity of reference station
4	✓	In coastal areas within beacon coverage	Within the European Maritime Area	Globally	In coastal areas within beacon coverage	In coastal areas within beacon coverage
5	✓	Within very restricted area	✓	✓	In coastal areas within a range of 100 km from reference station	In coastal areas within beacon coverage
6 (null group at present)	✓	✓	✓	✓	✓	Within close proximity of reference station
7	✓	In coastal areas within beacon coverage	Within the European Maritime Area	Globally	In coastal areas within beacon coverage	In coastal areas within beacon coverage
8	✓	Within very restricted area	✓	✓	In coastal areas within a range of 100 km from reference station	In coastal areas within beacon coverage
9	✓	✓	✓	✓	✓	Within close proximity of reference station
10 (null group at present)	✓	In coastal areas within beacon coverage	Within the European Maritime Area	Globally	In coastal areas within beacon coverage	In coastal areas within beacon coverage
11	✓	Within very restricted area	✓	✓	In coastal areas within a range of 100 km from reference station	In coastal areas within beacon coverage
12	✓	✓	✓	✓	✓	Within close proximity of reference station

29.3.2 Maritime Navigation

General navigation operations in the maritime sector as specified in Table 29.11 differentiates between five major phases: ocean and coastal waters, port approaches, restricted waters, and inland waterways. Safe and reliable navigation of vessels in each of these phases is based on GNSS and its augmentation systems. GNSS receivers used for navigation purposes vary significantly in terms of their positioning performance capabilities, robustness, and cost. GNSS receivers used for recreational and leisure navigation commonly use pseudorange measurements from both GPS and GLONASS satellites to determine the vessel position which drives many of the



Fig. 29.16 Trimble SPS351 Marine DGNSS/SBAS Receiver (courtesy of Trimble)

other features of the unit. Focusing primarily on plotting the position of the vessel on a basemap, features typically reflect those of in-car navigation systems such as route guidance and waypoint navigation as well as maritime specific details such as wind and tide information. More sophisticated models integrate other sensors such as 3-axis compass to provide heading information and a barometric altimeter to track pressure changes to determine elevation and to help monitor weather conditions.

Despite providing adequate positioning capabilities given the relatively low accuracy navigation or tracking requirements and the favorable satellite geometry conditions out on the open water, current generation GNSS faces a number of well-established shortcomings including a lack of basic integrity and limited accuracy without augmentation. Where higher levels of positioning accuracy and integrity are required, GNSS receivers like the Trimble SPS351 (Fig. 29.16), use the corrections and measurements provided by regional SBAS.

The beacon system of the International Association of Marine Aids to Navigation and Lighthouse Authorities (IALA) is the standard maritime GNSS augmentation system for many maritime applications. IALA GNSS beacons are used worldwide to provide DGNSS service for mariners [29.48]. For example, in the United States, the Coast Guard has successfully established a Nationwide Differential Global Positioning System (NDGPS) for maritime and land navigation. In Europe, beacons have been setup to provide DGNSS services across the European Maritime Area. Figure 29.17 is an

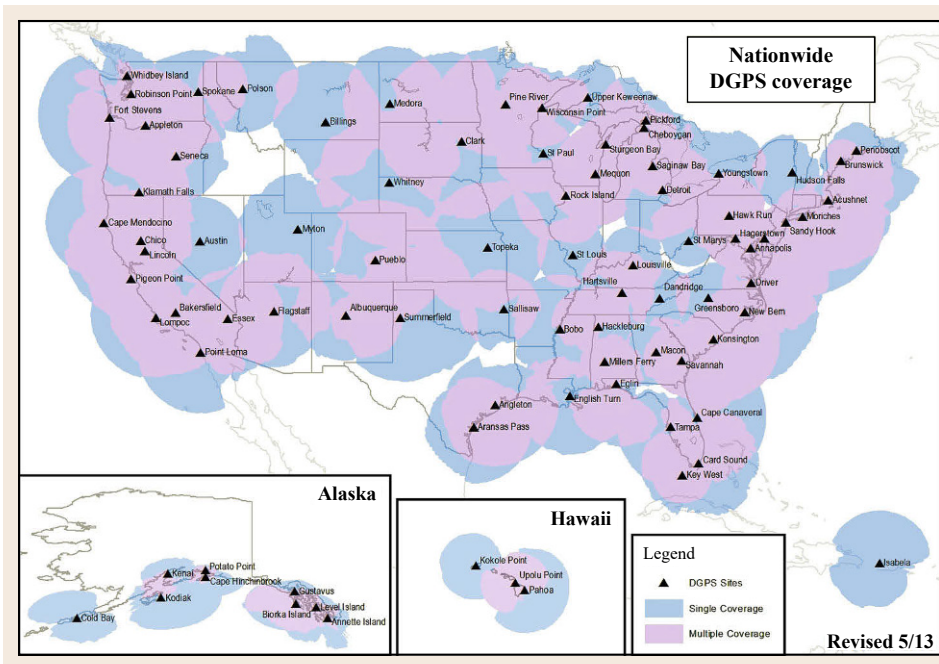


Fig. 29.17 United States' Differential GPS coverage (courtesy of United States Coast Guard)

example of the coverage of the differential network for the United States.

In Australia, DGPS corrections are broadcast from beacons setup along the coastline. Each DGPS beacon comprises two independent GPS receivers (for redundancy) and a radio transmitter operating in the LF/FM band 285–325 kHz. Typical range of coverage is approximately 150 NM. The corrections are broadcast in the Radio Technical Commission for Maritime Services (RTCM) standard (Annex A.1.3). Typical accuracies achievable with DGPS range between 2–4 m, with accuracies decreasing with range. Since the range of corrections is limited, DGPS using LM/FM radio is restricted to applications that are relatively close to the coastline. For vessels that need DGPS corrections in the open oceans a different method of receiving corrections is required. Receivers such as the Kongsberg DPS112 combined GPS L1/L2, GLONASS L1/L2, and SBAS receiver offer an integrated IALA beacon capability.

In addition, receivers of this class integrate the sub-meter accuracy from Fugro's Seastar SGG network of globally distributed dual system reference stations. These data are centrally processed to produce a global solution, under which corrections are calculated for each navigation satellite. These corrections are applied to the satellite ephemeris (orbit) and time reference clock information – hence, the term orbit and clock solution. This service utilizes the GPS L1 and L2 frequencies, and the GLONASS L1 and L2 frequencies, thereby providing an accurate measurement of variations in thickness of the ionosphere. This enables signal delay to be calculated and consequently a more accurate range/position is obtained. SGG provides a high availability, high integrity, global solution to a horizontal accuracy of better than 1 m (95%), and vertical accuracy of better than 1 m (95%).

DGNSS receivers typically use code signals to achieve a real-time accuracy of 1 m, 95%. For maritime navigation and operations, PPP techniques deliver decimeter level positioning accuracies using dual frequency carrier-phase measurements in combinations with real-time orbit and satellite clock corrections. The International GNSS Service (IGS) provides free access to real-time orbit and clock products using its global network of IGS reference stations, however, paid subscription services from Fugro's Starfire G4 and Veripos Ultra, provides these corrections using privately managed global networks of reference stations to compute and deliver these corrections via satellite communications. Starfire's service extends beyond GPS and GLONASS to include the BeiDou satellites as well. Enhanced-RTK like capabilities are achievable using Starfire G2+ service, which promises centimeter level

accuracies for GPS and GLONASS based on fixed integer ambiguities.

Furuno's SC-110 Satellite Compass (Fig. 29.18) determines a ship's heading by decoding the phase data in the GPS carrier frequency. In Fig. 29.19, a pair of antennas A1(ref) and A2(fore), each connected with an associated GPS engine and processor, are installed along the ship's fore-aft line. The GPS systems at A1 and A2 calculate the range and azimuth to the satellite. The difference in range between A1 and A2 is

$$\Delta\lambda + N\lambda,$$

where λ is the wavelength of the GPS L1 signal of 19 cm and N is the integer ambiguity resolved by the least-squares ambiguity decorrelation adjustment (LAMBDA) algorithm and automatically found during



Fig. 29.18
Furuno SC110
Satellite Compass
(courtesy of
Furuno)

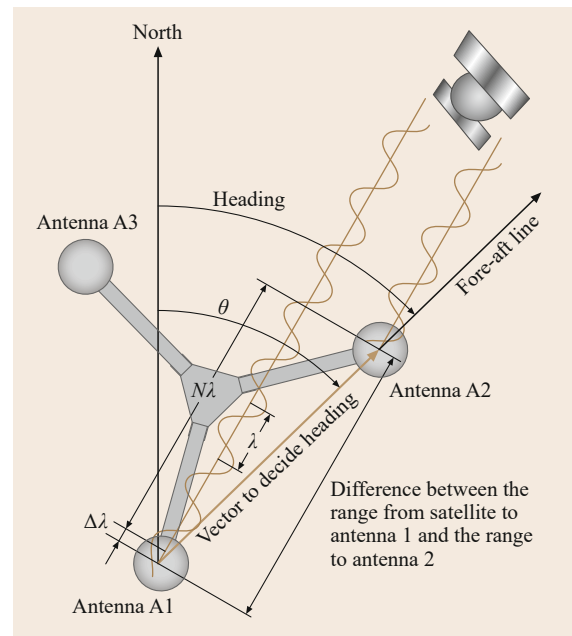


Fig. 29.19 Furuno SC110 Satellite Compass Computation
(courtesy of Furuno)

the initialization stage thus determining a vector (range and orientation) A1 to A2, i. e., heading of ship relative to north. In reality, a third antenna is added to reduce the influence of pitch, roll and yaw, and five satellites are used to process three-dimensional (3-D) data (by third sat), to reduce clock derived error (by fourth sat), and to calculate N in the initial stage (by fifth sat). If GPS signal is blocked by a tall building or the vessel is under a bridge, the 3-axis vibrating-gyro rate sensors in the processor unit take the place of the satellite until all five satellites are in view. The rate sensors also contribute to regulating the heading data against pitch, roll, and yaw together with the third antenna (A3 in Fig. 29.19).

29.3.3 eLoran

Enhanced long range navigation (eLoran) is an internationally standardized positioning, navigation, and timing (PNT) service for use by many modes of transport and in other applications. It is rapidly emerging as the primary GNSS backup for the IMO e-Navigation concept. eLoran is a low-frequency terrestrial navigation system based on a number of transmission stations, which emit precisely timed and shaped radio pulses centred at 100 kHz radio frequency. Each station emits a sequence of 8 pulses spaced 1000 ms apart. The stations are grouped into chains, which each consists of a single master station and two or more secondary stations. The master station transmits first, followed by successive transmissions from each of the secondary stations of the chain. The master/secondary transmission sequence is repeated periodically, with the period between repetitions called the Group Repetition Interval (GRI). eLoran is an independent, dissimilar com-

plement to GNSS. As such, it will allow PNT users with demanding safety-critical or mission-critical applications to secure their safety, security, and economic benefits even when their satellite services are disrupted. eLoran is capable of meeting the accuracy, availability, integrity, and continuity performance requirements for [29.49]:

- Aviation nonprecision instrument approaches
- Maritime harbor entrance and approach maneuvers
- Land-mobile vehicle navigation
- LBS
- Precise time and frequency users.

The concerns about the vulnerability of GNSS have sparked a renewed interest in the Loran PNT system. Recently, considerable effort has been put globally into investigating whether eLoran can provide a viable backup to GNSS [29.50, 51].

29.3.4 Automatic Identification System

The automatic identification system (AIS) is a maritime vessel location and tracking system designed to facilitate the automatic exchange of information between ships and between ships and shore-based authorities (vessel traffic services). The primary function of AIS is to enhance collision avoidance capabilities by providing individual ships with a picture of the marine traffic in their area. Consequently, IMO SOLAS regulations currently mandates AIS be fitted aboard international voyaging ships with gross tonnage of 300 or more, and all passenger ships regardless of size (Fig. 29.20).

AIS integrates a standardized very high frequency (VHF) transceiver (channels 161.975 and

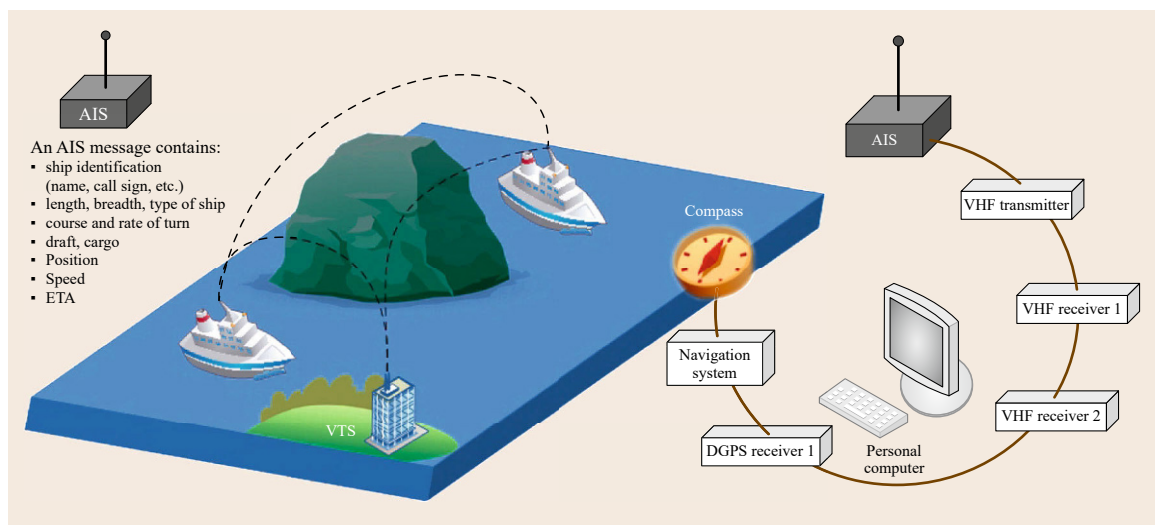


Fig. 29.20 AIS Architecture (courtesy of Shine Micro and US Coast Guard)

Table 29.15 Types and classes of AIS

AIS Class A	Class A has been mandated by the International Maritime Organization (IMO) for vessels of 300 gross tonnage and upwards engaged on international voyages, cargo ships of 500 gross tonnage and upwards not engaged on international voyages, as well as passenger ships (more than 12 passengers), irrespective of size.
AIS Class B	Class B provides limited functionality and is intended for non-SOLAS vessels. It is not mandated by the International Maritime Organization (IMO) and has been developed for non-SOLAS commercial and recreational vessels.
AIS Base Station	Base Stations are provided by an aids to navigation authorities to enable the ship to shore/shore to ship transmission of information. Networked AIS Base Stations can assist in providing overall maritime domain awareness.
AIS aids to navigation (AtoN)	AIS AtoN provide an opportunity to transmit position and status of buoys and lights through the same VDL, which can then show up on an electronic chart, computer display or compatible radar.
AIS SART	Search and Rescue Transmitters using AIS can be used to assist in determining the location of a vessel in distress.
AIS on Search and Rescue (SAR) Aircraft	Search and Rescue Aircraft may use AIS to assist in their operations.

162.025 MHz) with a positioning system such as a differential GNSS or LORAN-C receiver, and other electronic navigation sensors, such as a gyrocompass or rate of turn indicator. AIS transmits, automatically and at set intervals, dynamic information relating to the ship’s position, course, speed, and heading; static information related to the ships name, length, breadth; and voyage-related details such as cargo information and status (underway, at anchor). This data is sent every 2–10 s depending on a vessel’s speed while underway, and every 3 min while vessels are at anchor. AIS standard comprises several substandards called *types* that specify individual product types. The specification for each product type provides a detailed technical specification which ensures the overall integrity of the global AIS system within which all the product types must operate. The major product types described in the AIS system standards are: Class A and Class B, as well as different types of AIS used for shore stations (AIS Base Stations), aids to navigation (AIS AtoN), AIS on search-and-rescue (SAR) aircraft, and AIS SAR transmitters (AIS SART) (Table 29.15).

Vessels fitted with AIS transceivers and transponders can be tracked by AIS base stations located along coast lines or, when out of range of terrestrial networks, through a growing number of satellites that are fitted with special AIS receivers. Internet delivery of AIS information now provides the public with this information.

Collision Avoidance
AIS was developed by the IMO technical committees as a technology to avoid collisions among large vessels at sea that are not within range of shore-based systems. The technology identifies every vessel individually, along with its specific position and movements,

enabling a virtual picture to be created in real time. The AIS standards include a variety of automatic calculations based on these position reports such as Closest Point of Approach (CPA) and collision alarms. As AIS is not used by all vessels, AIS is usually used in conjunction with radar.

A vessel’s text-only AIS display lists nearby vessels’ range, bearings, and names. When a ship is navigating at sea, information about the movement and identity of other ships in the vicinity is critical for navigators to make decisions to avoid collision with other ships and dangers (shoal or rocks). Visual observation (e.g., unaided, binoculars, and night vision), audio exchanges (e.g., whistle, horns, and VHF radio), and radar or Automatic Radar Plotting Aid are historically used for this purpose. These preventative mechanisms, however, sometimes fail due to time delays, radar limitations, miscalculations, and display malfunctions and can result in a collision.

While the requirements of AIS are to display only very basic text information, the data obtained can be integrated with a graphical electronic chart or a radar display, providing consolidated navigational information on a single display.

Fishing Fleet Monitoring and Control
There is an increasing interest in ensuring that fishing and aquaculture resources are environmentally, economically, and socially sustainable. Strategies such as the the EU common fisheries policy are aimed at achieving this outcomes [29.52]. GNSS will play a significant role in monitoring and enforcing many regulatory requirements of this policy. Examples include, proving where fishing vessels are and managing fishing quotas. In a broader sense, GNSS will be central to any effort aimed at efficiently managing the conflict-

ing uses of maritime areas. Real-time and postprocessed GNSS positions will support many of the diverse activities included in maritime spatial planning (MSP), defined as [29.53]:

a process of analyzing and allocating parts of three-dimensional marine spaces to specific uses, to achieve ecological, economic, and social objectives that are usually specified through the political process.

AIS is widely used by national authorities to track and monitor the activities of their national fishing fleets. AIS enables authorities to reliably and cost effectively monitor fishing vessel activities along their coast line, typically out to a range of 60 miles (depending on location and quality of coast-based receivers/base stations) with supplementary data from satellite-based networks. The BDStar Navigation CDG-MF-08A BeiDou receiver (Fig. 29.21) is an advanced BeiDou receiver based on the technologies of positioning, communication, timing, and location-based information services of the BeiDou satellite. This receiver can manage a large amount of subordinated BeiDou subscriber machines in a wide range and monitor the ship position information and text message communication information of authorized subscribers in an area covered by the BeiDou system in real time, thereby protecting and conveying safety production in the marine fisheries.

Vessel Traffic Services

In busy waters and harbors, a local vessel traffic service (VTS) may exist to manage ship traffic. Here, AIS provides additional traffic awareness and information about the configuration and movements of ships.

Maritime Security

AIS enables authorities to identify specific vessels and their activity within or near a nation's Exclusive Economic Zone. When AIS data is fused with existing radar systems, authorities are able to differentiate between vessels more easily. AIS data can be automatically

processed to create normalized activity patterns for individual vessels, which when breached, create an alert, thus highlighting potential threats for more efficient use of security assets. AIS improves maritime domain awareness and allows for heightened security and control. Additionally, AIS can be applied to freshwater river systems and lakes.

Aids to Navigation

The AIS Aids to Navigation (AtoN) product standard was developed with the ability to broadcast the positions and names of objects other than vessels, such as navigational aid and marker positions and dynamic data reflecting the marker's environment (e.g., currents and climatic conditions). These aids can be located on shore, such as in a lighthouse, or on water, platforms, or buoys. The US Coast Guard has suggested that AIS might replace racon (radar beacons) currently used for electronic navigation aids [29.54].

AtoN's enable authorities to remotely monitor the status of a buoy, such as the status of the lantern, as well as transmit live data from sensors (such as weather and sea state) located on the buoy back to vessels fitted with AIS transceivers or local authorities. An AtoN will broadcast its position and identity along with all the other information. The AtoN standard also permits the transmit of *Virtual AtoN* positions whereby a single device may transmit messages with a *false* position such that an AtoN marker appears on electronic charts, although a physical AtoN may not be present at that location.

Search and Rescue

Deployed in 1982, the COSPAS-SARSAT is a worldwide satellite search-and-rescue (SAR) system that provides distress alert and location information to respective SAR authorities globally for maritime, aviation, and land users in distress [29.55]. SARSAT is an acronym for search and rescue satellite-aided tracking. COSPAS is an acronym for the Russian words *Cosmicheskaya Sistyema Poiska Avariynich Sudov*, which mean *Space System for the Search of Vessels in Distress*, indicative of the maritime origins of this distress alerting system.

The space segment for COSPAS-SARSAT consists of SAR payloads hosted by low-Earth orbiting (LEOSAR) and geostationary satellites (GEOSAR). To further improve the system performance, SAR capabilities will be hosted by the middle Earth orbits of the three GNSS constellations: GPS, Galileo, and GLONASS (MEOSAR). It is expected that by 2020 COSPAS-SARSAT will rely on a medium altitude Earth orbit *MEO/GEO* space segment, replacing the low Earth orbit *LEO/GEO* design [29.56]. With numerous



Fig. 29.21 BDStar Navigation CDG-MF-08A BeiDou Maritime Receiver (courtesy of BDStar)

satellites, each with an Earth coverage or footprint significantly larger than the LEO satellites (about seven times larger), the MEOSAR constellations will enable an instantaneous and worldwide coverage. Distress beacons will be detected and located more quickly and accurately than today, in as little as one beacon burst, that is, about 50 s. The more efficient alert notices that result will directly contribute to the efficiency of rescue operations where time is critical. A COSPAS-SARSAT beacon operating on 406 MHz frequency is used to indicate distress. There are three types of beacons: emergency locator transmitter (ELT) for aviation use, emergency position indicating radio beacon (EPIRB) for maritime use, and personal locator beacon (PLB) for general land or rail use. These beacons allow for digitally coded unique identification information in the beacon message, including location of the distress site based on GNSS (for new generation beacon models). In addition, second-generation beacons have a higher requirement for independent location accuracy.

For coordinating on-scene resources of a marine SAR operation, it is imperative to have data on the position and navigation status of other ships in the vicinity. In such cases, AIS can provide additional information and enhance awareness of available resources, even if the AIS range is limited to VHF radio range. The AIS standard also envisioned the possible use on SAR aircraft, and included a message (AIS Message 9) for aircraft to report their position.

To aid SAR vessels and aircraft in locating people in distress, the specification (IEC 61097-14 Ed 1.0) for an AIS-based SAR transmitter (AIS-SART) was developed by the IEC's TC80 AIS work group. AIS-SART was added to Global Maritime Distress Safety System regulations effective January 1, 2010. AIS-SARTs have been available on the market since at least 2009.

Future GNSS will contribute to the international SAR service, enhancing the worldwide performance of the current COSPAR-SARSAT system. The positioning accuracy of today's system is very poor (typically a few kilometers) and the alert is not always issued in real time. The Galileo SAR service will drastically reduce the time to alert, and the position of the distress beacon will be determined to within a few meters [29.57].

29.3.5 Shipping Container Tracking

The role of GNSS in automating commercial maritime activities has gained prominence given the significant operating efficiencies achievable. In many areas, it is the tracking capabilities of GNSS that can realize these benefits. In this section, the focus is on shipping con-

tainer tracking, identified as one of the major growth areas of maritime asset tracking, with millions of shipping containers flowing through ports all over the world on a daily basis [29.58]. This growth has been matched by developments in robust container tracking and tracing systems [29.59].

In [29.60] the Skema interactive knowledge based platform for transport and logistics highlight the use of GNSS for container tracking and tracing solutions. Differential and SBAS technologies for used for localization, combined with different telecommunication means (satellite and/or terrestrial) for communication. These combined GNSS-Telecommunication devices are installed on-board the containers/swap bodies/etc. Internationally, other forms of RTLS are used in ports. For example port of Singapore uses RFiD technology [29.61] with other proposed solutions based around wireless network technology and inertial sensors. In almost all cases, these solutions are integrated with the core GNSS capability [29.62].

While it is perhaps only necessary to locate an individual container to an accuracy of 10 m, operation in port areas and the task of placing containers onto cargo trucks and boats requires much finer levels of accuracy often at the millimeter level. Combined with associated problems of satellite visibility and multipath particularly in port areas, alternative positioning technologies have been readily integrated with GNSS as part of standard container tracking and gantry cranes.

Basic user requirements are for example [29.60]:

- Container tracking for the periodic real-time positioning of the swap body/container.
- Container integrity control.
- Geographical location related to automatic alerts and reports.
- Geofencing functions to raise a warning in case of deviation from a predefined route or in case of goods transiting into forbidden areas.
- Triggering of alarm with position data in case of tampering occurrence (e-seals anomalous data).
- Triggering of alarm in case the device does not show signs of life for a certain period of time indicating the possibility of tampering, removal, or destruction of the device.
- Times of departure, arrival, border crossing.
- Data management from inspections for statistical services.
- Operational history data logging and processing.
- Automatic and on-demand reports of the containers and their location and status.

29.4 Outlook

GNSS applications across the land, rail, and maritime sectors broadly reflect efforts to achieve multimodal C-ITS. As the whole of GNSS expands into these sectors, the challenges for meeting the performance requirements for safety and liability critical applications have necessitated the development of robust GNSS augmentation capabilities in terms of sensor fusion,

measurement integration algorithms and innovative use of current and emerging signals of opportunity. This chapter has presented the current status of some of these augmentation technologies and techniques, but it is without question that the future will see more applications and more approaches to delivering GNSS-like performance in the most difficult environments.

References

- 29.1 K. O'Keefe, S. Ryan, G. Lachapelle: Global availability and reliability assessment of the GPS and Galileo global navigation satellite systems, *Can. Aeronaut. Space J.* **48**(2), 123–132 (2002)
- 29.2 European GNSS Agency: *GNSS Market Report*, 4th edn. (Publications Office of the European Union, Luxembourg 2015)
- 29.3 *Global Navigation Space Systems: Reliance and Vulnerabilities* (The Royal Academy of Engineering, London 2011)
- 29.4 A. Rainio: Location-based services and personal navigation in mobile information society, *Int. Fed. Surv. (FIG) Working Week*, Seoul (FIG, Copenhagen 2001) pp. 1–14
- 29.5 S. Dhar, U. Varshney: Challenges and business models for mobile location-based services and advertising, *Commun. ACM* **54**(5), 121–128 (2011)
- 29.6 K.T. Dang, N.T. Phan, N.C. Ngo: An OpenLS privacy-aware middleware supporting location-based applications, *Int. J. Pervas. Comput. Commun.* **9**(4), 311–345 (2013)
- 29.7 S. Shek: Next-generation location-based services for mobile devices, *Lead. Edge Forum* (Computer Science Corporation, Falls Church 2010) p. 66
- 29.8 T. Reichenbacher: *Mobile Cartography: Adaptive Visualisation of Geographic Information on Mobile Devices* (Verlag Dr. Hut, Munich 2004)
- 29.9 NSTB/WAAS T&E Team: *GPS Performance Analysis Report #85* (William J. Hughes Technical Center, Atlanta City 2014)
- 29.10 EVA-7M u-blox 7 GNSS module, Data Sheet UBX 13000581-R07 (u-blox 2014)
- 29.11 A. Wieser, H. Hartinger: High-sensitivity GPS: Technologie und Anwendungen, *Proc. 66th DVW-Semin. GPS Galileo: Methoden, Lösungen neueste Entwickl.*, Darmstadt (DVW, Vogtsburg-Oberrotweil 2006) pp. 251–274, in German
- 29.12 F. van Diggelen: *A-GPS: Assisted GPS, GNSS, and SBAS* (Artech House, London 2009)
- 29.13 R. Mautz: *Indoor Positioning Technologies*, Habilitation Thesis (ETH, Zurich 2012)
- 29.14 R. Bill, C. Cap, M. Kofahl, T. Mundt: Indoor and outdoor positioning in mobile environments – A review and some investigations on wlan-positioning, *Geogr. Inf. Sci.* **10**(2), 91–98 (2009)
- 29.15 J. Cadman: Deploying commercial location-aware systems, *Proc. 2003 Workshop Locat.-Aware Com-*
- put. (held as part of UbiComp 2003), Seattle, ed. by M. Hazas, J. Scott, J. Krumm (2003) pp. 4–6
- 29.16 K. Siwiak, P. Withington, S. Phelan: Ultra-wide band radio: The emergence of an important new technology, *Proc. 53rd Veh. Technol. Conf. (VTC'2001)*, Rhodes Vol. 2 (2001) pp. 1169–1172
- 29.17 P.D. Groves: *Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems*, 2nd edn. (Artech House, London 2013)
- 29.18 A. Ledeczi, P. Volgyesi, J. Sallai, B. Kusy, X. Koutsoukos, M. Maroti: Towards precise indoor RF localization, *Proc. 5th Workshop Embed. Netw. Sens. (HotEmNets'08)*, Charlottesville (ACM, New York 2008) pp. 1–5
- 29.19 M. Bouet, A. L. Dos Santos: RFID tags: Positioning principles and localization techniques, *Proc. 1st IFIP Wirel. Days (WD'08)*, Dubai (2008) pp. 1–5
- 29.20 A. Lim, K. Zhang: A robust RFID-based method for precise indoor positioning, *Adv. Appl. Artif. Intell. 19th Int. Conf. Ind., Eng. Other Appl. Appl. Intell. Syst.*, IEA/AIE 2006, Annecy, ed. by M. Ali, R. Dapoigny (Springer, Berlin 2006) pp. 1189–1199
- 29.21 M.S. Grewal, L.R. Weill, A.P. Andrews: *Global Positioning Systems, Inertial Navigation, and Integration* (Wiley, Hoboken 2007)
- 29.22 A. Kealy, G. Roberts, G. Retscher: Evaluating the performance of low cost MEMS inertial sensors for seamless indoor/outdoor navigation, *Proc. IEEE/ION PLANS 2010*, Indian Wells (2010) pp. 157–167
- 29.23 M. Efatmaneshnik, N. Alam, A.T. Balaei, A. Kealy, A.G. Dempster: Cooperative positioning in vehicular networks. In: *Wireless Technologies in Vehicular Ad Hoc Networks: Present and Future Challenges* (IGI Global, Mexico City 2012) pp. 245–270
- 29.24 ARRB Project Team: *Vehicle Positioning for C-ITS in Australia* (Research Report AP-R431-13, Austroads 2013)
- 29.25 J.P. Tripathi: Algorithm for Detection of Hot Spots of Traffic Through Analysis of GPS Data, Ph.D. Thesis (Thapar University, Patiala 2010)
- 29.26 J. Beugin, J. Marais: Simulation-based evaluation of dependability and safety properties of satellite technologies for railway localization, *Transp. Res. C Emerg. Technol.* **22**, 42–57 (2012)
- 29.27 GNSS Rail User Forum: Requirements of Rail Applications (European GNSS Secretariat, Brussels 2000)

- 29.28 Standard for communications-based train control (CBTC) performance and functional requirements, IEEE Std 1474.1-2004 (2004)
- 29.29 A. Mirabadi, N. Mort, F. Schmid: Application of sensor fusion to railway systems, Proc. IEEE/SICE/RSJ Int. Conf. Multisens. Fusion Integr. Intell. Syst., Washington (1996) pp. 185–192
- 29.30 A.C. Knight, H. Uebel: System for indicating track sections in an interlocking area as occupied or unoccupied, US Patent 4 763 267 (1988), Alcatel N.U.
- 29.31 G. Barbu: GNSS/GALILEO certification for rail safety applications railway requirements and the strategic position of UIC, Proc. 8th World Congr. Railw. Res. (WCRR'2008), Seoul (UIC, Paris 2008) pp. 1–9
- 29.32 K.-H. Shin, D. Shin, E.-J. Eui-Jin Joung, Y.-G. Kim: The reliability and safety enhancement method of GNSS for train control application, Proc. 23rd Int. Tech. Conf. Circuits/Syst., Comput. Commun. (ITC-CSCC 2008), Shimomoseki (UIC, Paris 2008) pp. 1545–1548
- 29.33 E. González, C. Prados, V. Antón, B. Kennes: GRAIL-2: Enhanced odometry based on GNSS, Procedia – Soc. Behav. Sci. **48**, 880–887 (2012)
- 29.34 B. Allotta, V. Colla, M. Malvezzi: Train position and speed estimation using wheel velocity measurements, Proc. Inst. Mech. Eng, F J. Rail Rapid Transit **216**(3), 207–225 (2002)
- 29.35 S. Bedrich, X. Gu: GNSS-based sensor fusion for safety-critical applications in rail traffic, Proc. NAVITEC'2004, Noordwijk (ESA, Netherlands 2004) pp. 1–8
- 29.36 F. Senesi: Satellite application for train control systems: The test site in Sardinia, J. Rail Transp. Plan. Manag. **2**(4), 73–78 (2012)
- 29.37 F. Rispoli, A. Filip, M. Castorina, G. Di Mambro, A. Neri, F. Senesi: Recent progress in application of GNSS and advanced communications for railway signaling, Proc. 23rd Int. Conf. Radioelektron., Pardubice (2013) pp. 13–22
- 29.38 K. Williams, M.J. Olsen, G.V. Roe, C. Glennie: Synthesis of transportation applications of mobile LIDAR, Remote Sens. **5**(9), 4652–4692 (2013)
- 29.39 L.-S. Tey, L. Ferreira, H. Dia: Evaluating cost-effective railway level crossing protection system, Proc. 32nd Australas. Transp. Res. Forum, Auckland (ACT, Canberra 2009) pp. 1–12
- 29.40 J. Greenbaum: *Real-Time Asset Management for Railroad Freight: The RfTrax Opportunity* (Enterprise Applications Consulting, Berkley 2006)
- 29.41 A. Rosová, M. Balog, Ž. Šimeková: The use of the RFID in rail freight transport in the world as one of the new technologies of identification and communication, Acta Montan. Slovaca **18**(1), 26–32 (2013)
- 29.42 V. Scinteie: Implementing passenger information, entertainment, and security systems in light rail transit, 9th Natl. Light Rail Transit Conf., Portland (Transp. Res. Board, Transp. Res. Circ. E-C058, Washington 2003) (UIC, Paris 2003) pp. 528–533
- 29.43 P. Parker: Real-time information: Need for, reliability and management, <http://melbourneontransit.blogspot.com.au/2010/08/real-time-passenger-information-need.html>, last accessed 21 January 2014
- 29.44 A. Grant, P.L. Williams, N.K. Ward, S. Basker: GPS jamming and the impact on maritime navigation, J. Navig. **62**(2), 173–187 (2009)
- 29.45 Z. Kopacz, W. Morgas, J. Urbanski: The changes in maritime navigation and the competences of navigators, J. Navig. **57**(1), 73–83 (2004)
- 29.46 M. Fairbanks, N. Ward, W. Roberts, M. Dumville, V. Ashkenazie: GNSS augmentation systems in the maritime sector, Proc. ION NTM 2004, San Diego (ION, Virginia 2004) pp. 662–673
- 29.47 C. Dixon, R.G. Morrison: A pseudolite-based maritime navigation system: Concept through to Demonstration, J. Global Position. Syst. **7**(1), 9–17 (2008)
- 29.48 G. Mangs, S. Mittal, T. Stansell: Worldwide beacon DGPS status and operational issues, RTCM Orlando, FL (Leica Geosystems, Torrance 1999)
- 29.49 S. Basker, P. Williams: Navigating eLoran: Challenges and the way forward, Proc. XVIIth IALA Conf. (2010)
- 29.50 D. Last: GNSS: The present imperfect, Inside GNSS **5**(3), 60–64 (2010)
- 29.51 G.W. Johnson, P.F. Swaszek, R.J. Hartnett, R. Shalae, M. Wiggins: An evaluation of eLoran as a backup to GPS, Proc. 2007 IEEE Conf. Technol. Homel. Secur. (IEEE, 2007) pp. 95–100
- 29.52 H. Frost, P. Andersen: The common fisheries policy of the European Union and fisheries economics, Marine Policy **30**(6), 737–746 (2006)
- 29.53 C. Ehler, F. Douvere: Marine spatial planning, a step-by-step approach towards ecosystem-based management. Intergovernmental Oceanographic Commission and Man and the Biosphere Programme. IOC Manual and Guides No. 53, ICAM Dossier No. 6 (Intergovernmental Oceanographic Commission and Man and the Biosphere Programme, UNESCO, Paris 2009)
- 29.54 I. Harre: AIS adding new quality to VTS systems, J. Navig. **53**(3), 527–539 (2000)
- 29.55 J.V. King: Overview of the Cospas-Sarsat satellite system for search and rescue, Proc. 6th Int. Mob. Satell. Conf. (IMSC'99), Ottawa (Communications Research Center, Nepean 1999)
- 29.56 S.D. Ilcev: Development of Cospas-Sarsat satellite distress and safety systems (SDSS) for maritime and other mobile applications. In: *Marine Navigation and Safety of Sea Transportation: Navigational Problems*, ed. by A. Weintritt (CRC Press, London 2013) p. 269
- 29.57 A. Lewandowski, B. Niehoefer, C. Wietfeld: Performance evaluation of satellite-based search and rescue services: Galileo vs. COSPAS-SARSAT, Proc. IEEE 68th Vehic. Technol. Conf. 2008, VTC 2008–Fall (IEEE, 2008) pp. 1–5
- 29.58 J.M. Moreno: Bar seal for shipping container, US Patent 7 044 512 (2006)
- 29.59 W.K. Talley: Ocean container shipping: Impacts of a technological improvement, J. Econ. Issues **34**(4), 933–948 (2000)

- 29.60 G. Lynch: e-Maritime Overview, SKEMA Sustainable Knowledge Platform for the European Maritime and Logistics Industry, SST-2007-TREN-1 – SST.2007.2.2.4., Feb 2010
- 29.61 R. Angeles: RFID technologies: Supply-chain applications and implementation issues, *Inf. Syst. Manag.* **22**(1), 51–65 (2005)
- 29.62 H.K. Maheshwari, A.H. Kemp, Q. Zeng: Range based real time localization in wireless sensor networks. In: *Wireless Networks, Information Processing and Systems*, ed. by D.M.A. Hussain, A.Q.K. Rajput, B.S. Chowdhry, Q. Gee (Springer, Berlin 2009) pp. 422–432

Aviation App

30. Aviation Applications

Richard Farnworth

The Global Positioning System (GPS) has been available for civilian use for the past three decades and is now extensively used in aviation to support multiple applications.

This chapter describes how GNSS is used in aviation, the performance requirements that are being applied and the operational applications that have been enabled. It describes how conventional navigation has been gradually replaced by area navigation, the global introduction of Performance Based Navigation (PBN) and how the availability of GNSS has played a significant role in that evolution. The performance requirements for the different phases of flight are presented including the different methods by which the navigation integrity is ensured.

The goal of this chapter is to provide the reader with an overview of how GNSS has been adopted in aviation and explain how it has been integrated onto the aircraft alongside other navigation systems. The regulatory and certification process is also described to introduce the mechanisms by which aircraft operators can get approval to use GNSS in their daily operations.

30.1	Overview	878	30.3	Evolution of the Flight Deck	884
30.1.1	Conventional Navigation	878	30.3.1	The Navigation Data Chain	885
30.1.2	Area Navigation – RNAV	879	30.3.2	General Aviation	885
30.1.3	The Arrival of GNSS	880	30.3.3	Helicopters	885
30.2	Standardising GNSS for Aviation	881	30.4	From the RNP Concept to PBN	886
30.2.1	Aircraft Based Augmentation Systems	882	30.4.1	Performance Based Navigation (PBN)	886
30.2.2	Satellite Based Augmentation Systems	882	30.4.2	Navigation Specifications	886
30.2.3	Ground Based Augmentation Systems	883	30.5	GNSS Performance Requirements	888
			30.5.1	Description of the Relevant Parameters	888
			30.5.2	GNSS Integrity Concepts	890
			30.6	Linking the PBN Requirements and the GNSS Requirements	891
			30.6.1	Phases of Flight	891
			30.6.2	RNAV Approaches	893
			30.6.3	RNP AR APCH	895
			30.7	Flight Planning and NOTAMs	897
			30.8	Regulation and Certification	897
			30.8.1	Airworthiness Certification	897
			30.8.2	Operational Approvals	898
			30.9	Military Aviation Applications	898
			30.10	Other Aviation Applications of GNSS	899
			30.10.1	Surveillance (ADS-B)	899
			30.10.2	Datalink	899
			30.11	Future Evolution	900
			30.11.1	GNSS Vulnerability and Alternative-PNT	900
			30.11.2	Rationalisation of the Navigation Infrastructure	900
			30.11.3	Multi-Constellation	901
			References		901

30.1 Overview

The aviation community has embraced the use of GNSS as it provides a high performance global navigation capability meeting many of the requirements of aviation users. Furthermore, the cost, size and flexibility of GNSS receivers make it an attractive navigation solution for all airspace users. The global drive towards performance based navigation is creating an even stronger role for GNSS and together with external augmentation systems is enabling the introduction of new kinds of approach procedures which can improve safety and provide better access to airports. GNSS will be the cornerstone for many aviation applications in the future. However, there are still concerns with the single frequency, single constellation civil aviation uses today. The main issue is vulnerability linked to the robustness of the signals and the potential for electromagnetic interference; both unintentional and intentional. A multi-constellation/multi-frequency (MC/MF) environment should alleviate the majority of these issues but will introduce additional complexity. In order to explain its role and the way aviation uses GNSS this chapter will describe the evolution from conventional navigation to today's PBN environment and show how the different aviation applications have benefitted from the introduction of GNSS.

30.1.1 Conventional Navigation

Instrument navigation evolved from simple radio direction finding. By the 1920s aviation specific navigation systems had been developed such as radio ranging otherwise called A/N transmitters. This was based on a network of radio transmitters broadcasting the Morse code letters *A* and *N*. The pilot would listen to the broadcast tones and when the aircraft was on course he would hear a steady tone. When left of track he would hear the Morse code letter *A* and when right of track he would hear *N*. This system, which by the 1930s had defined a series of air routes across the continental US (CONUS), was the forerunner of the Non Directional Beacon (NDB). Today, the NDB is considered the most basic navigation aid; a simple radio transmitter which radiates a single frequency omnidirectional signal. Although the system is over seventy years old new NDB transmitters are still being installed today. In the aircraft, the receiving antenna was originally manually adjusted to find the signal bearing but very quickly became automated and is today called Automatic Direction Finding (ADF). This navigation capability allows the pilot to identify the relative bearing to the beacon. One of the challenges for the aircrew using NDB was that to *home-in* to a station the pilot

had to correct for wind, so flying a specific bearing to/from the station required some skill. The introduction of VHF Omnidirectional Range (VOR) in the 1950s provided a specific bearing, or radial, to the ground aid but with significantly enhanced accuracy compared to the NDB. Figure 30.1 illustrates an aircraft flying on a radial towards a VOR station marked by the small hexagon symbol. It also illustrates a position *fix* at the intersection of radials from two VOR stations. Aircraft navigation was further enhanced in the 1960s by the introduction of Distance Measuring Equipment (DME) which, as the name suggests, provides range information. DMEs are often co-located with VOR stations to provide the pilot not only a radial but also a distance to the navigation aid.

As air transport grew in the 1930s and 40s the recognition that formal coordination of traffic flows and the management of the airspace was necessary. From the late 1940s until the late 1990s the structure of the airspace and the Air Traffic Services (ATS) routes was designed around the location of ground-based navigation aids. The ATS routes were point-to-point connecting one ground aid to another. When

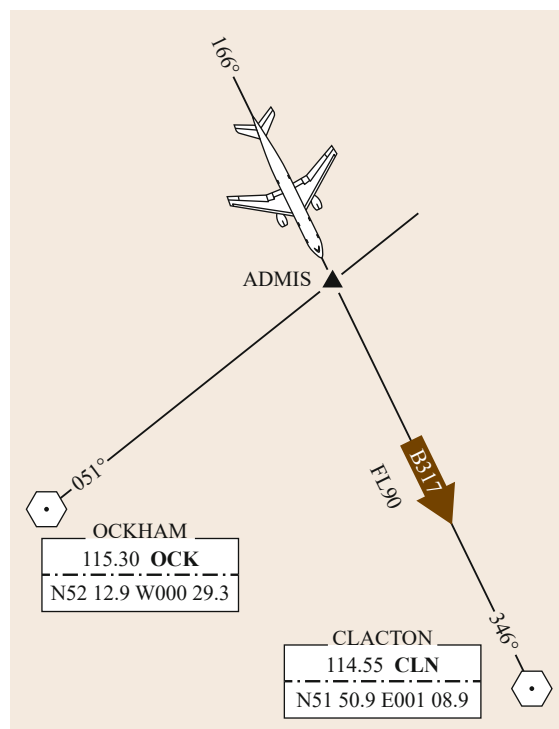


Fig. 30.1 Conventional navigation: the aircraft is approaching a position fix, ADMIS, identified by the intersection of two VOR radials

new ATS routes were required additional navigation aids had to be installed to support them. If location or cost did not permit a new navigation aid then a *fix* could be defined provided the supporting navigation aids were within range. A *fix* is the point defined by the intersection of two radials or a range and bearing from a collocated VOR/DME. Where airports did not have precision instrument landing systems these same navigation aids were used to provide guidance for Non-Precision Approach (NPA) procedures. The location of the navigation aids was often therefore a compromise between the en-route and approach navigation requirements.

As the continental ATS route structure was defined by tracks to and from the ground navigation aids the pilots normally flew them by manually selecting aircraft heading, observing the bearing and distance to the aid and correcting the track to take account of the wind. The resulting navigation accuracy achieved was around 5 nmi (nautical miles). This is the 2-sigma value indicating that aircraft will be within 5 nmi of the desired track for at least 95% of the flight time. However, performance is not uniform as the precision of the VOR improves the closer to the station the aircraft is. The safety and protection requirements of the ATS routes were based on this expected performance. One major disadvantage of conventional navigation was that as the traffic increased and more ATS routes were implemented to manage the aircraft, more and more routes were starting or terminating at each navigation aid. Bottlenecks were appearing in the system. In the 1990s traffic saturation started to appear in certain places of the European network and this was the driver for action.

30.1.2 Area Navigation – RNAV

In the late 1970s the first digital avionics started appearing on civil aircraft. The Lockheed L1011 Tristar is credited as the first commercial aircraft with a navigation computer. The computer was called Carousel and it allowed pilots to manually enter coordinates and then use inertial navigation to guide the aircraft to that position. The pilots started using these early computers to fly the conventional ATS routes by inserting the coordinates of the ground based aids into the computer and allowing the navigation system to provide the guidance along the required inbound radial. These early computers were the forerunner of the modern, complex Flight Management System (FMS) which looks after the navigation, guidance and many other features of the aircraft. The ability to fly to a coordinate rather than to the overhead of a navigation aid provided flexibility and led to the development of Area Navigation (RNAV) which is defined as follows [30.1]:

A method of navigation which permits aircraft operations on any desired flight path within the coverage of ground or space-based navigation aids or within the limits of the capability of self-contained aids, or a combination of these.

Figure 30.2 illustrates an RNAV route between a series of waypoints superimposed on a conventional route which passes over the groundbased navigation aids. The onboard computer used the same navigation aids to compute its position but the pilot was no longer obliged to fly from one aid to the next.

The evolution of on-board navigation computers which provided an area navigation capability enabled aircraft to be operated on any track between two geographic points-in-space independent of the location of the ground navigation aids. Aircraft now had the ability to fly to waypoints as well as tracking bearings between ground aids. This, theoretically at least, allowed new routes to be designed anywhere without requiring additional navigation aids to be installed. RNAV was available in the 1970s and 80s using VOR and DME inputs.

Despite these advances the route structure in air traffic systems remained firmly based on the location of the ground navigation aids. With only a partial equipage of the aircraft fleet the route structure could not be adapted. Handling a mix of aircraft, some with RNAV capability and some without, and the assignment by ATC of RNAV or conventional routes depending on aircraft capability proved to be unworkable. To overcome this problem in Europe a mandate for the carriage of *Basic-RNAV* equipment [30.2] finally came into force

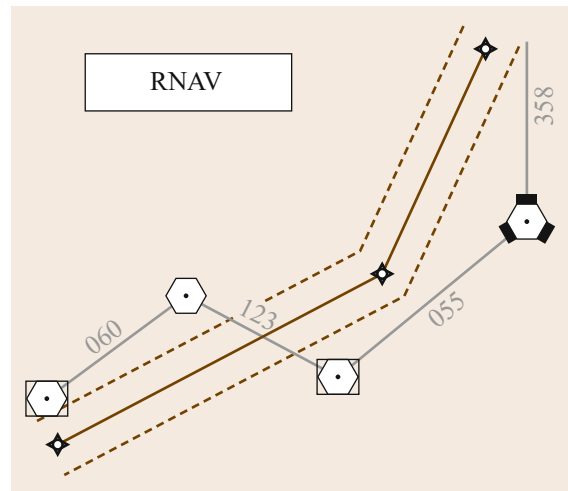


Fig. 30.2 Conventional and area navigation between waypoints independent from the location of the ground based aids

in April 1998 more than twenty years after commercial aircraft with an area navigation capability had started rolling off the production line.

30.1.3 The Arrival of GNSS

At the same time as the aviation community was realising the potential of area navigation and addressing how to gain benefit from the capability of RNAV systems, the American Navstar Global Positioning System (GPS) was evolving. GPS could support RNAV systems globally so coverage of navigation aids would no longer be an issue. The GPS programme had been initiated in 1973 and by 1978 the first satellite was in orbit. During the 1980s the majority of the constellation was deployed and although it was not fully operational, GPS played an important role in the first Gulf War. Following the loss of flight KAL007, a Korean Airlines Boeing 747, which was shot down by the Soviets in 1983, President Reagan authorised that civilians should be given access to GPS. However, when the civilians were granted access to the GPS signal, the US military did not want their positioning system to be used against them and so they deliberately degraded the performance of the Standard Positioning Service (SPS) by a slight jittering of the atomic clock and a slight corruption of the navigation message. This deliberate degradation in performance on the primary frequency, L1, was called Selective Availability (SA) [30.3, 4]. Even with this known degraded performance, the aviation community could see the potential of this positioning system which provided global coverage and a position accuracy that was far better than any existing conventional ground-based navigation aid [30.5].

However, aviation is a highly regulated industry and if GPS was to be used on-board aircraft there was a need to develop international global standards. The avionics industry, the Radio Technical Commission for Aeronautics (RTCA), produced the first industry standard for GPS receivers in 1991 [30.6] and it was entitled *Minimum Operational Performance Standards for Airborne Supplemental Navigation Equipment Using Global Positioning System* (RTCA DO-208). This standard was soon followed by a European equivalent (ED-72-A) issued by Eurocae, the European Organisation for Civil Aviation Equipment and the US Federal Aviation Administration (FAA) Technical Standard Order, TSO C-129 [30.7], which provided the means by which GPS equipment could be installed and integrated on-board commercial aircraft and certified for operational use. It should be noted from the title that DO-208 and its European equivalent ED-72A allowed only for *supplemental* use of GPS at that time. This meant that the GPS equipment did not replace the conventional

navigation equipment on the aircraft but was carried in addition and that the ground-based navigation aids needed to be kept operational in case of problems with GPS. This requirement is clearly demonstrated in the Joint Aviation Authorities (JAA) certification document for basic area navigation [30.8] which allowed the use of GPS as a RNAV sensor but with limitations.

The introduction of GPS as a navigation system was relatively clear cut for the aircraft manufactures, avionics manufacturers and even to the aircraft operators. A GPS receiver was another navigation box that had to be integrated and certified like any other on-board system. However, for the wider aviation community the GPS signal-in-space (SIS) was an enormous change and this was particularly true for the national air navigation service providers who provided Air Traffic Control and the enabling communications, navigation and surveillance services.

International Standardisation

International standards for civil aviation are governed by the International Civil Aviation Organisation (ICAO), a United Nations agency, and laid down in the Convention on International Civil Aviation, also known as the Chicago Convention, which was signed in 1944 by the initial 52 contracting States. Over the last 60 years, the ICAO family has grown and today has 191 signatory States. ICAO develops global Standards and Recommended Practices (SARPS) which are binding on the Member States unless they file for non-compliance and Procedures and Air Navigation Services (PANS) which may or may not be applied in each Member State.

Under Article 28 of the convention Contracting States are required to:

undertake, so far as it may find practicable, to provide, in their territory [...] air navigation facilities to facilitate international air navigation, in accordance with the standards and practices recommended or established from time to time, pursuant with the Convention.

Until the advent of global positioning systems, every State had always provided their own navigation facilities within their territories. The arrival of global positioning systems brought a dilemma for some States. Here was a system providing global coverage but under single State control and in the hands of the military as well! The natural question that arose was *What responsibilities would a State be assuming by allowing the use of such a system for air navigation within its airspace?* [30.9]: It has taken some States a long time to overcome this institutional hurdle, there has been much debate and not all of them have yet succeeded

in allowing GPS to be used for anything other than supplemental navigation. The lack of control over core constellations GPS and GLONASS, both under military control, and the unwillingness to become dependent on signals provided by a single foreign State has been partly the driver for the development of the new and evolving satellite positioning systems such as Europe's Galileo, the Chinese BeiDou (also known as Compass) and the Indian Regional Navigation Satellite System (IRNSS/NavIC).

In 1994 the US Government offered, in a letter from the FAA Administrator to ICAO, that the:

GPS standard positioning service was to be made available on a continuous worldwide basis and free of direct user fees for the foreseeable future. At least 6 years notice prior to termination.

A similar offer concerning GLONASS was made in 1996 by the Russian Federation.

A common Geodetic Reference System

The introduction of area navigation, which is based on the ability of the aircraft's computer to fly between coordinates, highlighted the need for the entire aviation community to move towards a common geodetic

reference system rather than the national systems historically used. The introduction of GPS with its own global reference system reinforced the need for a uniformly coordinated reference system and the aviation community accepted this change. The ICAO Assembly agreed in 1989 to adopt WGS-84 [30.10], the latest evolution of the militarily developed World Geodetic System, as the standard geodetic system to be applied globally. With every State using their own national reference system, each developed separately for national needs and in the majority of cases not being compatible with each other; a coordinate given in one system did not bring the user to the same position in a neighbouring States coordinate system. As all reference systems are mathematical models, conversion from one model to another is possible. However, with the global implementation of WGS-84 in the 1990s it became very apparent that transformation from the national datum to WGS-84 was not a simple task [30.11] as in many cases the original survey data was either not available or of insufficient accuracy to support the precision needed for the aviation RNAV requirements. Therefore, the WGS-84 implementation programme took many years and even today some States have filed *non compliance* against this ICAO standard.

30.2 Standardising GNSS for Aviation

In 1993, ICAO established the GNSS Panel and tasked it with the objective to develop Standards and Recommended Practices (SARPs) for the use of GNSS in aviation. GNSS was given the following broad definition [30.12]:

A worldwide position and time determination system that includes one or more satellite constellations, aircraft receivers and system integrity monitoring, augmented as necessary to support the required navigation performance for the intended operation.

The way GNSS was standardised in ICAO was a significant departure from the way standardisation had been done for other, traditional ground-based navigation systems. The purpose of standardising a conventional navigation aid such as a DME was to ensure interoperability. A DME transponder (the ground station) could be produced by a variety of global manufacturers and each manufacturer's transponder would need to behave exactly the same way in response to an airborne DME interrogator. With GPS the constellation was already global and the signals were defined by the US military in the Interface Control Document (ICD) 200 series and

even though other GNSS systems would be developed later there would not be one constellation or augmentation with exactly the same signal characteristics as the other. The purpose of developing the SARPs for GNSS was to ensure that the aviation community had a common understanding of the parameters of the signal that could be assumed by the manufacturers of aviation receivers. Some States argued that the GPS signal was already defined in the ICD and the Standard Positioning Service (SPS) Performance Standard published by the US government [30.13]. These States felt it was not beneficial to reproduce all or part of the two documents in an ICAO standard. Eventually, it was decided to develop the ICAO GNSS standards [30.14] and they were published in Annex 10 [30.12] in 1999 and became globally applicable in 2001.

For civil aviation, whilst the accuracy of the position estimation is important, the confidence that the estimation is not corrupted is equally as important. This level of confidence is called *integrity* [30.15] and civil aviation has set a 10^{-7} requirement on the positioning system to alert the crew if there is a problem; that equals one missed detection in 10 000 000. Although the US DoD had considered a unique integrity channel when

GPS was being developed in the 1970/80s, it was never implemented due to cost. Therefore, for civil aviation integrity needed to be *added* to make the signal acceptable and this addition is referred to as augmentation.

ICAO has standardised three different types of augmentation system for GNSS as follows:

- Aircraft Based Augmentation System – **ABAS**
- Satellite Based Augmentation System – **SBAS**
- Ground Based Augmentation System – **GBAS**.

These systems are shortly described below. For detailed information on SBAS and GBAS, readers are referred to Chaps. 12 and 31, respectively.

30.2.1 Aircraft Based Augmentation Systems

This type of augmentation is the simplest form of integrity monitoring and can be segregated into two techniques. The first is to integrate the GPS receiver with inputs from other on-board sensors such as accurate clocks, altimetry systems or inertial platforms. This form of ABAS was termed Aircraft Autonomous Integrity Monitoring (AAIM). The second technique, which is the most widely applied form of ABAS used with GPS, is Receiver Autonomous Integrity Monitoring (RAIM). RAIM is an algorithm within the GPS receiver which makes use of redundant signals from the satellite constellation due to the availability of more than the minimum set of satellites needed to compute a position. A receiver requires four satellite measurements to compute a 3-dimensional position solution and its time. As the constellation was originally designed with 24 satellites and the actual number of satellites in the constellation has been consistently above that, more than the minimum number are generally visible

to the user receiver. With those extra satellites available the algorithm in the user receiver can check the relative consistency of the different measurements and identify if a faulty satellite exists. This is called fault detection and a warning can be issued to the pilot if the receiver detects a fault. A warning is also given if the number of satellites is not sufficient to perform the RAIM check. If six or more satellites are visible some RAIM algorithms are capable of not only detecting the fault but also identifying which is the erroneous ranging source. This enables the receiver to exclude the faulty satellite and continue to provide a position with integrity; this functionality is called Fault Detection and Exclusion (FDE). As the ICAO standards are intended to define the Signal-in-Space and RAIM is a function implemented in the user equipment, the aviation specifications for RAIM can be found in the avionics equipment standards developed by organisations such as the RTCA and Eurocae [30.6].

30.2.2 Satellite Based Augmentation Systems

SBAS uses a regional network of ground monitoring stations to assess the core constellation and provide a navigation message to users through Geostationary Satellites. The SBAS system architecture is shown in Fig. 30.3. The navigation message created by the SBAS master control station provides range corrections and integrity information for each satellite in view of the ground monitoring network. If a GPS satellite correction becomes too large, then a *Do-Not-Use* flag will be sent for that satellite. These corrections are applied in the user equipment and provide both improved positioning accuracy and integrity.

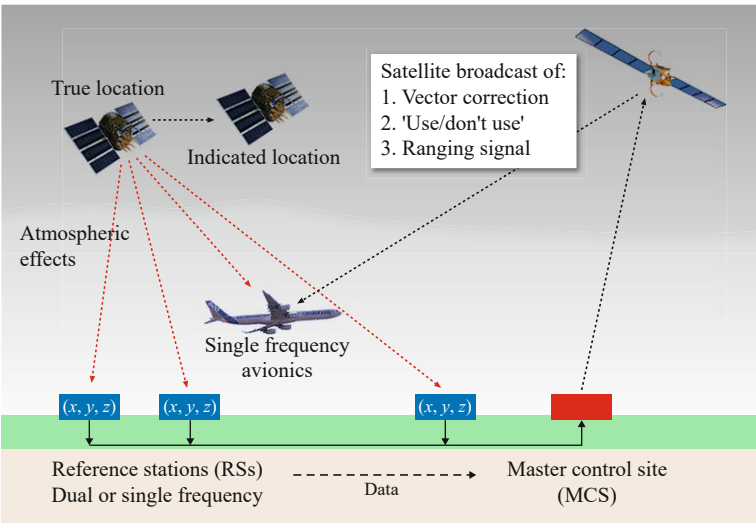


Fig. 30.3 SBAS system architecture. Correction messages are broadcast to the user through a geostationary satellite

Several SBAS systems have been implemented or are under development in different regions of the world beginning with the US Wide Area Augmentation System (WAAS) [30.16], and followed by the European Geostationary Navigation Overlay Service (EGNOS) [30.17], the Japanese Multi-functional Satellite Augmentation Service (MSAS) [30.18], the Indian GPS Aided Geo Augmented Navigation (GAGAN) [30.19] and the emerging Russian System for Differential Corrections and Monitoring (SDCM) [30.20]. The reader should note that currently only SDCM provides monitoring of GPS and GLONASS. All the other SBAS systems only augment GPS. SBAS can provide older and less capable aircraft with a high end navigation capability with relatively little investment.

30.2.3 Ground Based Augmentation Systems

This augmentation system provides local monitoring of the core constellation and works on the principle that the error at the location of the ground station is the same as that of the aircraft. Whilst this is true when the aircraft is close to the ground station, the differences increase when the aircraft is further away. The GBAS system architecture is shown in Fig. 30.4. GBAS uses a series of very accurately located antennas on or around an aerodrome which feed the received signals to the processing unit allowing the system to calculate the time delay on each signal in view of the station. The processing unit generates a series of navigation messages which are transmitted to approaching aircraft using a VHF data broadcast (VDB) transmitter. The system also holds a series of unique approach paths in the approach database and upon selection by the ground station operator, the GBAS will transmit the desired final approach segment (FAS) for the aircraft to follow.

GBAS systems are intended to support precision approach and landing operations and ultimately replace the conventional Instrument Landing System (ILS).

A major benefit of GBAS over traditional precision approach and landing systems (ILS and MLS) is that a single station can support multiple runway ends and in theory, at least, can support multiple aerodromes which are in close proximity to each other. For ILS and MLS each runway requires a dedicated system and antenna installed close to the runway to support precision approach and landing operations. Therefore, GBAS has the potential to offer service providers with a cost effective solution to support multiple runway ends. Furthermore, as GBAS does not have to be located at the runway end, airports which previously could not support a precision approach due to geographical limitations and antenna siting issues may now be able to provide that capability. Another important benefit of GBAS over ILS is the potential to increase runway throughput during low visibility operations. When Low Visibility Procedures are implemented the ILS systems require a greater protection of critical and sensitive areas to ensure the radio signals emitted from the ground transmitters located close to the runway are not interfered with by any moving object such as an aircraft. Therefore, in low visibility conditions extra spacing between aircraft on approach is required. This can have a significant impact on the runway throughput. As GBAS does not have the same critical and sensitive areas as ILS the spacing between aircraft can be maintained in all weather conditions.

There are three different categories of ILS which support three different categories of approach and landing operations. These approaches bring aircraft to a Decision Height (DH) above the runway where the pilot must have sufficient visual references to be able to complete the landing or initiate a missed approach

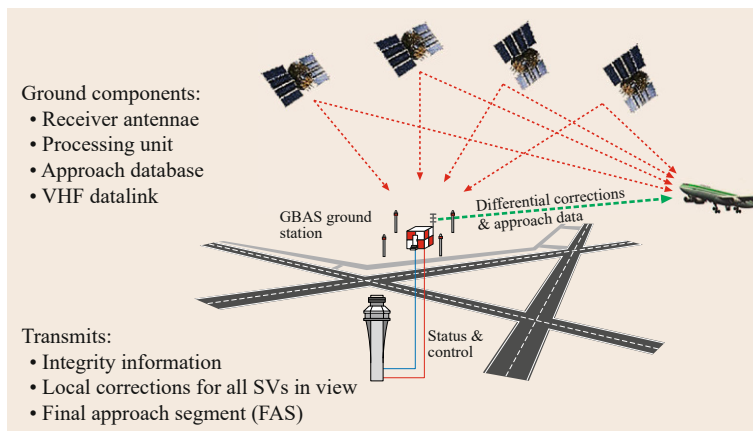


Fig. 30.4 GBAS system architecture

procedure:

- Category I: Minimum DH = 200 ft
- Category II: Minimum DH = 100 ft
- Category III: No DH.

30.3 Evolution of the Flight Deck

Up until the 1970s, flight crews played an active role in the generation of the position solution and the utilisation of the navigation information to meet the airspace requirements. The pilot followed information provided on printed charts showing the routes and the location of the ground based navigation aids. Pilots manually tuned the required navigation aids, identified the station and then used the signal to locate the aircraft on the ATS routes. With the advent of digital avionics this role started to change. From as early as the 1980s Flight Management Systems (FMS) started to include the ability to perform automatic navigation. Even with added automation the flight crew were still required to monitor the aircraft trajectories with reference to conventional navigation aids as the FMS was only approved for what was called supplemental navigation as mentioned earlier. In the 1990s along with the arrival of GPS the systems became more integrated and a monitoring function was added where the actual navigation performance being achieved was displayed to the pilot and could be checked against the required navigation performance for the route. This led to a more hands-off approach for the flight crew as the routes were programmed into the FMS and workload was considerably reduced.

Commercial aircraft use multi-function displays such as can be seen in Fig. 30.5. The displays are fed



Fig. 30.5 The cockpit of a commercial aircraft with Flight Management System (courtesy of Airbus)

GBAS Category I systems are already operational in the US, Europe, Australia and Russia and development is ongoing to enable these systems to evolve to support Category II and III operations in the future.

by the flight management computer and the various avionics equipment that are located in the avionics bay with all the other boxes that are Line Replaceable Units (LRUs) with a standardised form and interface. A typical GNSS avionics receiver used in commercial aircraft is shown in Fig. 30.6.

However, flight management systems became extremely complex and their development was not harmonised. Systems evolved in different ways as the different FMS manufacturers implemented a wide range of functionalities [30.21]. Issues became apparent due to the different performance of various aircraft with different FMS types or even different software versions of the same manufacturers FMS. The published ATS routes were not flown in exactly the same way, the aircraft did not always turn at the same point and the accuracy of the trajectory was not harmonised. This issue of mixed performance, commonly referred to as *mixed mode* operations, had to be addressed. The computer systems relied on input data which was formalised in an industry standard document published by Aeronautical Radio Incorporated, ARINC 424 [30.22], describing the navigation system database which became the reference document for interchange of data between different airborne systems. As navigation was becoming more automated and the navigation performance relied on high quality data, the responsibility of the providers of the navigation database increased and detailed processes were needed to ensure that data integrity could be guaranteed.



Fig. 30.6 The GNSS receiver of a commercial aircraft to be installed in the avionics bay (courtesy of Rockwell Collins)

30.3.1 The Navigation Data Chain

Each country is required by ICAO to provide an Aeronautical Information Publication (AIP) which contains all the details of regulations, procedures and other information necessary for flying in that country's airspace. Data houses use the information published in the AIPs to create the source navigation data in ARINC 424 format that will be used by the FMS suppliers to create the on-board navigation databases.

Because of the importance in the correctness of the navigation data the whole process is regulated by a number of standards beginning with the ICAO Annex 15 which establishes global standards for aeronautical information services (AIS). Industry standards covering data source providers and FMS database providers [30.23, 24] have been developed to establish clear processes to make sure that the navigation data quality is assured throughout the data chain. Today the data suppliers are audited against the industry standards for the processing of aeronautical data by the regulatory authorities in an effort to ensure high quality data.

Changes are continuously being made in the aviation environment due to the introduction of new routes, navigation aids and procedures. Therefore a fixed 28 day amendment cycle for the on-board navigation database has been defined within ICAO Annex 15 and this is called the Aeronautical Information Regulation and Control (AIRAC) cycle. The AIRAC cycle, first adopted in 1964, requires that the aeronautical databases are always up to date. It is essential, for both efficiency and safety, that Pilots, Air Traffic Controllers, Air Traffic Flow Managers, Flight Management Systems and Aviation Charts all have the same data set. The AIRAC cycle defines effective dates which are globally harmonised to ensure that every stakeholder in the system applies changes at the same time.

30.3.2 General Aviation

Due to their cost and complexity flight management systems were originally only available on large commercial air transport type aircraft. Pioneers in the development of General Aviation RNAV systems began to appear in the 1980s based on Loran-C positioning inputs. These units provided the basic RNAV functions to provide guidance along a route between a series of waypoints. When GPS started becoming available at the beginning of the 1990s these same RNAV functionalities were re-engineered around a GPS navigation engine which replaced the Loran-C element for providing the position information. Garmin, the market leader in general aviation GNSS navigation units has built on this experience to create equipment available for general

aviation aircraft that has almost all of the same navigation functionalities as the FMS used in commercial airliners [30.25]. A typical in general aviation navigation unit is shown in Fig. 30.7. This is a stand-alone unit installed in the cockpit instrument panel.

One of the issues that arose in the early days was the need for appropriate pilot training. General aviation pilots now had an on-board computer with hundreds of different functions and it was essential that they knew how to use and monitor these functions correctly.

30.3.3 Helicopters

Helicopter operations are often performed in challenging environments such as mountainous areas for emergency medical services, flights to offshore oil platforms and at low level in built up urban environments. When visibility is poor their operations are severely limited. The use of GNSS technology has enabled new procedures to be introduced that enhance safety and allow flights in conditions where visibility is limited and there is no surveillance available. Operations to offshore oil platforms are often outside the coverage of ground based navigation aids and in bad weather the use of



Fig. 30.7 A typical general aviation GNSS navigation unit (courtesy of Garmin)

weather radar was very common as a means of navigating towards the platforms and avoiding obstacles. The weather radar was never designed for this purpose and safety concerns were raised by regulatory authorities. The introduction of GNSS based RNAV procedures has significantly contributed to safety improvements for

helicopter operations. Dedicated instrument flight procedures for helicopters allow for Point-in-Space (PinS) designs and support approaches to hospital landing platforms. SBAS LPV procedures have allowed operations to lower operational minima than was ever possible previously.

30.4 From the RNP Concept to PBN

The early use of area navigation systems had simply automated the following of the conventional route structure and terminal procedures. GPS was not introducing new applications but providing an alternative means of doing what had been done before, albeit with much greater accuracy.

In the early 1990s ICAO published the RNP Concept which told the world what lateral navigation performance was required to fly on a particular route but did not prescribe how to achieve it. As the decade progressed the increased traffic demand and the pressure for greater efficiency led to area navigation being used more broadly and new routes were introduced that could only be flown by RNAV capable aircraft. However, due to the lack of harmonized standards different RNAV solutions led to different area navigation applications being implemented in different regions of the world [30.26]. Operators needed to get approval against different specifications and provide specific crew training to perform basically the same operation and this was becoming unnecessarily complex and costly. With the proliferation of different terminology there was also a lot of room for confusion.

To address this ICAO established the Required Navigation Performance and Special Operational Requirements Study Group (RNPSORSG) in 2004 with the aim of harmonising the use of area navigation globally and to standardise the terminology being used. This led to the development of the Performance Based Navigation (PBN) concept which was published in the ICAO PBN Manual [30.1] in 2007.

30.4.1 Performance Based Navigation (PBN)

The PBN concept, which superseded the RNP concept, builds on the older concept by not only stipulating the required performance but also how it can be achieved. PBN has three components as illustrated in Fig. 30.8:

- A Navigation Infrastructure
- A Navigation Specification

which together support a

- Navigation Application.

The PBN manual [30.1] specifies RNAV system performance requirements independently from the navigation sensor. It also includes the functionalities needed for particular applications. It provides this information in a series of 11 navigation specifications that are defined at a sufficient level of detail to facilitate global harmonization [30.27].

PBN takes account of the fact that RNAV systems have developed over a 40 year period and as a result there are a large variety of implementations (Fig. 30.9). Identifying navigation requirements, rather than the means of meeting the requirements, allows the use of all RNAV systems meeting these requirements irrespective of the means by which they are met. Technologies can evolve over time without requiring the operation itself to be revisited, as long as the requisite performance is provided by the RNAV system.

30.4.2 Navigation Specifications

A Navigation Specification is a set of aircraft and air crew requirements needed to support a navigation application within a defined airspace concept. The Navigation Specification defines the performance required of the RNAV system as well as any functional requirements such as the ability to conduct curved path procedures or to fly parallel offset routes. The Air Nav-

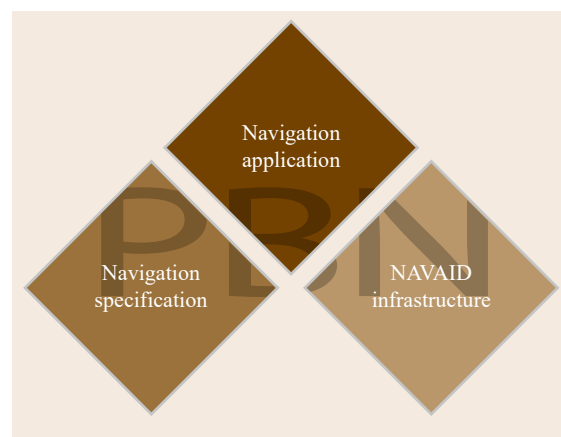


Fig. 30.8 The three components of the PBN concept

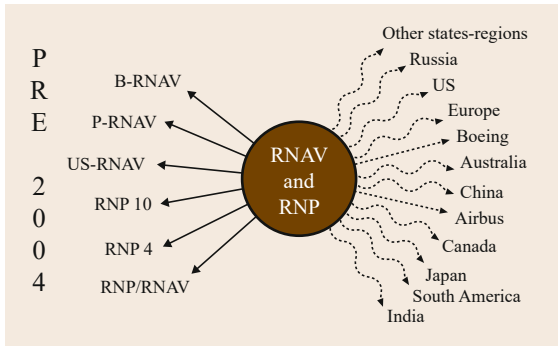


Fig. 30.9 The proliferation of navigation standards around the world

igation Service Provider (ANSP) must also ensure that the available navigation aid infrastructure (both ground and/or space based) can support the users in meeting the requirements.

PBN assumes that future navigation (with the exception of Precision Approach which is not addressed in the PBN Manual) will be undertaken by the use of an Area Navigation system, either standalone or, incorporated into a flight management system which provides performance management capability (3D and ultimately 4D navigation). It was identified that with this greater reliance on the RNAV system a higher level of integrity would be needed as the pilot will no longer be in a position to cross check the navigation performance using conventional aids. To provide such increased integrity there would need to be an indicator to the pilot of the achieved performance so that he can confirm that the requirements for the airspace are continuously being met. The PBN concept therefore includes navigation specifications that have a requirement for on-board performance monitoring and alerting

(OPMA). Navigation specifications that include a requirement for OPMA are *RNP specifications* and those not requiring OPMA are *RNAV specifications*. The RNAV specifications are the legacy operations. All the new navigation specifications have RNP in the title as they require OPMA.

GNSS, which includes OPMA by default with its integrity monitoring on the position solution is an ideal contributor to the achievement of RNP, although additional means are required on the aircraft to monitor the flight technical error. This effectively results in all RNP specifications requiring GPS as the positioning sensor.

Table 30.1 summarises the specifications included in the ICAO PBN manual [30.1]. The title of the navigation specification generally begins with either RNAV or RNP followed by a number which is the navigation accuracy requirement. This is the total system accuracy required to be met by the aircraft including both the error of the navigation system and the error linked to the ability of the pilot to accurately follow the desired path. The Advanced RNP specification does not include a number in the title as it is an umbrella specification including a variety of navigation accuracy requirements supporting different phases of flight. The specifications supporting the approach phase of flight are named RNP APCH and RNP AR APCH where AR is an abbreviation for *Approval Required* indicating that operators need specific operation approvals in order to fly such procedures.

To support these different specifications, the accuracy of the position estimation is key. Table 30.2 identifies which navigation sensors support which specification in the PBN manual. It can be seen that GNSS is a sensor that supports all of the specifications. The table also identifies which specifications require the aircraft to have an Automatic Flight Control System (AFCS) which may be Autopilot (AP) or Flight Director (FD).

Table 30.1 PBN navigation specifications by phase of flight. The numbers in each column represent the 95% accuracy requirement in nautical miles for the phase of flight. Where there is no number it means that the navigation specification is not suitable for this phase of flight

Navigation specification	En route oceanic/remote	En route continental	Arrival	Flight phase				Departure
				Initial	Intermediate	Final	Missed	
RNAV 10	10							
RNAV 5		5	5					
RNAV 2		2	2					2
RNAV 1		1	1	1	1		1	1
RNP 4	4							
RNP 2	2	2						
RNP 1			1	1	1		1	1
Advanced RNP	2	2 or 1	1	1	1	0.3	1	1
RNP APCH				1	1	0.3	1	
RNP AR APCH				1–0.1	1–0.1	0.3–0.1	1–0.1	
RNP 0.3		0.3	0.3	0.3	0.3		0.3	0.3

Table 30.2 Navigation sensors and AFCS requirements by navigation specification

	GNSS	IRU	Permitted sensors			AFCS requirement AP/FD
			DME/DME	DME/DME/IRU	DME/VOR	
RNAV 10	✓	✓				FTE may be manually controlled by the pilot remaining within 1/2 full scale deflection of CDI with correct scaling for phase of flight
RNAV 5	✓	✓	✓	✓	✓	
RNAV 1/2	✓		✓	✓		
RNP 4	✓					
RNP 2	✓					
RNP 1	✓		✓			✓
A-RNP	✓		✓			
RNP 0.3	✓					✓
RNP APCH	✓		✓ ^a	✓ ^a		
RNP AR APCH	✓					✓

^a For the initial and intermediate phases only

Most specifications do not require AP/FD and it is expected that flying manually the pilot will remain within half of the full scale deflection of the Course Deviation indicator (CDI).

It should be noted that although DME/DME can support the accuracy required for some RNP navigation

specification this is highly dependent on the availability and the geometry of the DME ground stations. In addition, due to different aircraft architectures some aircraft cannot provide On-board Performance Monitoring and Alerting (OPMA) with DME/DME.

30.5 GNSS Performance Requirements

ICAO requirements for GNSS signals-in-space are contained in Annex 10 to the convention on Civil Aviation [30.12]. The table of requirements from this ICAO Annex is reproduced in Table 30.3. There are different requirements depending on the operation to be performed but they all use the same four parameters of Accuracy, Integrity, Continuity of Service and Availability. Note the use of the term GNSS which indicates that these are the navigation system performance requirements for any GNSS system, not just GPS.

The requirements are presented as signal-in-space requirements. However, some of the parameters, such as positioning accuracy, cannot be applied as signal-in-space requirements without defining the user receiver. A GNSS position solution requires the combination of signals from several satellites. As a result, ICAO developed the concept of a fault-free receiver with a defined performance to be used to measure the signal-in-space. The accuracy and time-to-alert requirements are applicable to the navigation system performance at the output of the user receiver. The integrity risk, continuity and availability are applicable to the signal-in-space.

30.5.1 Description of the Relevant Parameters

Accuracy is a measure of the position error, which is the difference between the estimated position and

the true position, that will be experienced by a user with a certain probability at any instant in time. In general, the probability used for the accuracy requirements in Table 30.3 is 95%. For example the en-route horizontal accuracy requirement of 3.7 km means that under nominal circumstances there is a 95% probability that the horizontal position error is < 3.7 km.

Integrity is a measure of the trust that can be placed in the correctness of the position solution. Integrity includes the ability of a system to provide timely and valid warnings (alerts) to the user. More specifically the integrity risk is defined as the probability that a user will experience a horizontal position error larger than the horizontal alert limit (HAL) or a vertical position error larger than the vertical alert limit (VAL) without an alert being raised within the specified Time-to-alert (TTA) at any instant in time. Integrity is the important parameter that provides monitoring of the position solution and is a safety driver for aviation users. The integrity mechanism uses alert limits which are set at a certain probability level. The alert limit for GNSS systems used in aviation is generally set to a probability of 10^{-7} . That means that the probability of a position error exceeding the alert limit without a warning to the user is 10^{-7} . In Table 30.3 the integrity requirement is written as $1-10^{-7}$ which means that the probabil-

Table 30.3 The ICAO GNSS Signal-in-Space Performance Requirements (after [30.12], courtesy of ICAO)

Typical operation	Accuracy horizontal ^{a,c}	Accuracy vertical ^{a,c}	Integrity ^b	Time-to-alert ^c	Continuity ^d	Availability ^e
En-route	3.7 km	N/A	$1-1 \cdot 10^{-7}/h$	5 min	$1-1 \cdot 10^{-4}$ to $1-1 \cdot 10^{-8}/h$	0.99–0.99999
Terminal	0.74 km	N/A	$1-1 \cdot 10^{-7}/h$	15 s	$1-1 \cdot 10^{-4}$ to $1-1 \cdot 10^{-8}/h$	0.99–0.99999
Non-precision approach (NPA)	220 m	N/A	$1-1 \cdot 10^{-7}/h$	10 s	$1-1 \cdot 10^{-4}$ to $1-1 \cdot 10^{-8}/h$	0.99–0.99999
APV-I ^h	16 m	20 m	$1-2 \cdot 10^{-7}$ in any approach	10 s	$1-8 \cdot 10^{-6}$ per 15 s	0.99–0.99999
APV-II ^h	16 m	8 m	$1-2 \cdot 10^{-7}$ in any approach	6 s	$1-8 \cdot 10^{-6}$ per 15 s	0.99–0.99999
Category 1 precision approach ^g	16 m	6–4 m ^f	$1-2 \cdot 10^{-7}$ in any approach	6 s	$1-8 \cdot 10^{-6}$ per 15 s	0.99–0.99999

^a The 95th percentile values for GNSS position errors are those required for the intended operation at the lowest height above threshold (HAT), if applicable.

^b The definition of the integrity requirement includes an alert limit against which the requirement can be assessed. For Category 1 precision approach, a vertical alert limit (VAL) greater than 10 m for a specific system design may only be used if a system-specific safety analysis has been completed. The alert limits are:

Typical operation	Horizontal alert limit	Vertical alert limit
En-route (oceanic/continental low density)	7.4 km (4 nmi)	N/A
En-route (continental)	3.7 km (2 nmi)	N/A
En-route (terminal)	1.85 km (1 nmi)	N/A
NPA	556 m (0.3 nmi)	N/A
APV-I	40 m (130 ft)	50 m (164 ft)
APV-II	40 m (130 ft)	20 m (66 ft)
Category I precision approach	40 m (130 ft)	35–10 m (115–33 ft)

^c Accuracy and time-to alert requirements include the nominal performance of a fault-free receiver.

^d Ranges of values are given for the continuity requirement for en-route, terminal, initial approach, NPA and departure operations, as this requirement is dependent on several factors including the intended operation, traffic density, complexity of airspace and availability of alternative navigation aids. The lower value given is the minimum requirement for areas with low traffic density and airspace complexity. The higher value given is appropriate for areas with high traffic density and airspace complexity. Continuity requirements for Approaches with Vertical Guidance (APVs) and Category I operations apply to the average risk (over time) of loss of service, normalized to a 15 s exposure time.

^e A range of values is given for the availability requirements as these requirements are dependent upon the operational need which is based upon several factors including the frequency of operations, weather environments, the size and duration of the outages, availability of alternate navigation aids, radar coverage, traffic density and reversionary operational procedures. The lower values given are the minimum availabilities for which a system is considered to be practical but are not adequate to replace non-GNSS navigation aids. For en-route navigation, the higher values given are adequate for GNSS to be the only navigation aid provided in an area. For approach and departure, the higher values given are based upon the availability requirements at airports with a large amount of traffic assuming that operations to or from multiple runways are affected but reversionary operational procedures ensure the safety of the operation.

^f A range of values is specified for Category I precision approach. The 4.0 m (13 ft) requirement is based upon ILS specifications and represents a conservative derivation from these specifications.

^g GNSS performance requirements for Category II and III precision approach operations are under review and will be included at a later date.

^h The terms APV-I and APV-II refer to two levels of GNSS approach and landing operations with vertical guidance (APV) and these terms are not necessarily intended to be used operationally.

ity of the position error being smaller than the alert limit is 99.99999%.

Figure 30.10 illustrates the accuracy requirement, given at the 95% or 2σ level and the alert limit, assuming a Normal distribution of errors set at 5.3σ .

Algorithms in the user equipment compute a protection level each time a position solution is calculated. If the protection level exceeds the required alert limit for the operation then an alert is raised within the time-to-alert.

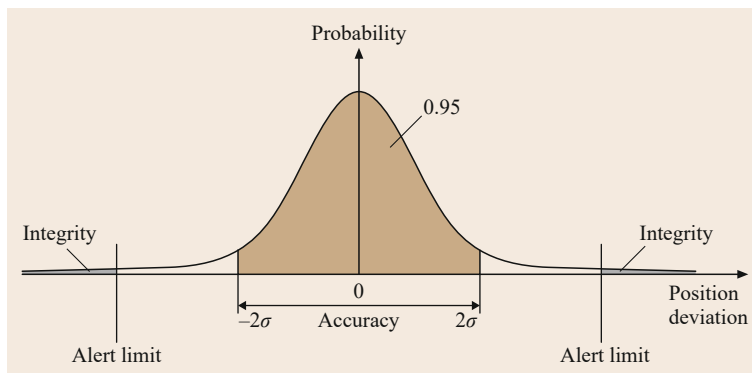


Fig. 30.10 Integrity, alert limits and accuracy

Continuity is defined as the probability that a user is able to determine its position with the specified accuracy and is able to monitor the integrity of its determined position over the time interval applicable for the corresponding phase of flight [30.28]. Assuming the service is available at the start of an operation this is the probability of it becoming unavailable over a specified time interval that is linked to the duration of the operation. Taking, for example, the en-route continuity requirement and choosing the lower value from the range of $1-10^{-4}/h$. If the service is available at the start of the hour then the probability of losing the service in the following hour is 10^{-4} . For approach operations which are short in duration the time period over which the continuity requirement is specified is 15 s.

Availability is defined as the probability that a user is able to determine its position with the specified accuracy and is able to monitor the integrity of its determined position at the initiation of the intended operation. Again there is a range of values in the ICAO table and a particular value of requirement needs to be selected for a particular operation depending on various factors described in table note ^c to the Table [30.29]. Assuming the lower value is taken of 0.99 this means that when a user expects to start an operation then the probability of them having a position meeting the accuracy requirement with integrity is 99%.

30.5.2 GNSS Integrity Concepts

Receiver Autonomous Integrity Monitoring (RAIM)

The basic integrity mechanism applied in aviation GNSS user equipment is RAIM [30.30–33]. RAIM uses redundant information from the GNSS constellation itself to verify integrity. As previously stated there needs to be more than the minimum number of satel-

lites needed to compute a position so that the receiver can check the consistency between the signals from different satellites. The basic RAIM algorithms only perform fault detection. When a fault is detected the GNSS system can no longer be used for navigation. More advanced RAIM algorithms perform Fault Detection and Exclusion (FDE) where a faulty satellite can be removed and a position solution with integrity can still be provided to the user. Multiple satellite failures will probably not be detected by such RAIM algorithms.

Satellite Based Augmentation Systems (SBAS)

The integrity algorithms used in an SBAS receiver make use of signals derived from a network of ground monitoring stations that provide error correction and quality information for each of the satellite signals. The navigation message containing this information is broadcast to users over a wide area using geostationary satellites. The signal is broadcast on the GPS L1 frequency and can also be used as an additional ranging source. The receiver applies integrity algorithms that have been standardized in [30.34] to compute the vertical and horizontal protection levels. The SBAS integrity concept is described in detail in Chap. 12. SBAS enables approach procedures with minima equivalent to ILS category I.

GBAS

The integrity concept for GBAS is more complex but applies the same principles to provide a series of parameters, based on local ground measurements, to be applied by the user equipment to compute the protection levels. The GBAS navigation message is broadcast to the aircraft using a VHF data-link. The GBAS integrity concept is described in detail in Chap. 31.

GBAS is intended as a precision approach and landing aid to replace today's ILS for all categories of approach and landing up to Category III.

30.6 Linking the PBN Requirements and the GNSS Requirements

The navigation specifications in the ICAO PBN Manual [30.1] define the performance and functionalities required to support different aviation applications. The GNSS requirements in ICAO Annex 10 [30.12] are applicable only to the navigation system. The difference is most apparent in the accuracy requirements as illustrated in Fig. 30.11.

The RNP lateral accuracy defined within the navigation specification is the Total System Error (TSE), to which the aircraft's performance will be certified. The TSE is made up of three error components (Fig. 30.11):

- **Path Definition Error (PDE)** – is the difference between the desired path that the designer wishes the aircraft to fly over the ground and the path that is calculated by the navigation computer. The PDE is assumed to be very small (almost negligible) and is tightly controlled by the oversight of the navigation data chain.
- **Navigation System Error (NSE)** – is the difference between the true position and the estimated position. This error is influenced by which navigation sensor is providing the position estimation.
- **Flight Technical Error (FTE)** – is the difference between the estimated position and the defined path. It is a measure of how well the pilot or the avionics can follow the guidance information provided by the navigation system.

When using GNSS as a position sensor the accuracy is well within the requirements for most phases of flight and the TSE, which is the difference between the track the aircraft should be following and the one it is actually following, will be dominated by the FTE.

30.6.1 Phases of Flight

The navigation requirements vary with the different phases of flight. Table 30.1 shows the navigation specifications applicable to each phase of flight. The first

column in the GNSS requirements, Table 30.3, lists typical operations and there is not a direct correlation between the two. This can lead to confusion.

En route – Oceanic and Remote

In Oceanic and remote areas where terrestrial navigation aids cannot be installed, the route separation was originally based on inertial navigation systems and their specified drift rates which are measured in nautical miles per hour. This is the phase of flight where GNSS had the first significant impacts in aviation. FAA notice 8110.57, issued in 1995, authorised the use of GPS equipment meeting specific requirements to be used as a primary means of navigation for oceanic operations. This effectively allowed aircraft that were not fitted with inertial navigation systems to fly the routes over the ocean using GPS equipment. There had to be a dual receiver installation with RAIM FDE and a RAIM Availability prediction was required to be performed as part of the flight preparation.

It can be seen from Table 30.1 that several different navigation specifications can be applied in oceanic airspace. The typical one in use today is RNAV-10, but future plans are to apply RNP 4 or RNP 2 in some oceanic regions and therefore allow route spacing to be reduced. The minimum spacing between routes in RNAV-10 airspace is 50 nmi. When using GNSS as the navigation sensor in oceanic regions, it must meet the requirements of the first line of Table 30.3. Positioning accuracy of 3.7 km (2 nmi) is easily achievable and is far better than the total system accuracy requirement of 10 nmi 95%. This leaves plenty of margin to accommodate the FTE performance.

The continuity requirement in the RNAV-10 navigation specification is set at $10^{-5}/h$ with an alert limit for integrity set at 20 nmi. Alert limits in GNSS receivers are generally set to the value of 4 or 2 nmi, miles which are sufficiently large to ensure availability requirements are met. The continuity requirement is driven by failure rates of hardware and this figure tends

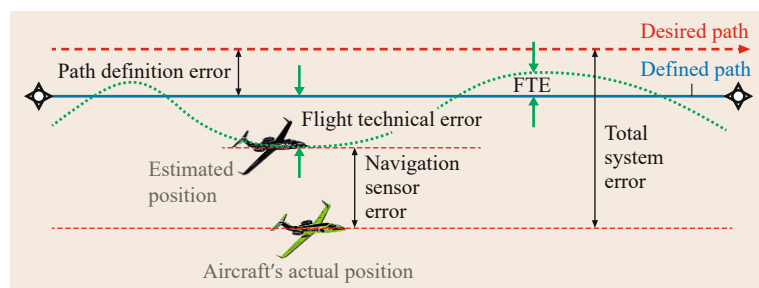


Fig. 30.11 Total system error components

to lead to a requirement for dual equipment to ensure redundancy.

In this phase of flight there is generally no surveillance and Air Traffic Control is based on procedural separation rules with pilots sending position reports every 10° of longitude to the ATC centre via an HF radio link. The ability to automatically provide position reports through a satellite data link is becoming more the norm in oceanic regions enabling lower lateral and longitudinal separation minima to be applied. A more advanced system of Space-based Automatic Dependent Surveillance Broadcast (ADS-B) using satellite communication services is now under development for use in oceanic airspace (Sect. 30.10.1).

En-route – Continental

The en-route phase is between the end of a departure procedure and the beginning of an arrival procedure. This is most of the core continental airspace where in busy regions of the world there is full coverage of surveillance radar and ground based navigation aids such as VOR and DME. Aircraft navigate along predefined routes which require a certain navigation performance from the aircraft. The European route network has a mandate for RNAV5 capability with a lateral accuracy requirement for the Total System Error (TSE) of ± 5 nmi. Other regions have similar requirements although the accuracy requirement varies between 1 nmi and 5 nmi. The route spacing in an RNAV 5 environment can be between 18 and 10 nmi depending on the ATC intervention capability.

GNSS is not the only sensor that can be used to support RNAV-5 and the requirements on continuity, integrity and availability in the navigation specification are applicable to all possible sensors. For GNSS the requirements are in the first line of Table 30.3. A 2 nmi alert limit is used even though the performance required from the navigation sensor is 2 nmi 95%.

Terminal Area – Arrivals and Departures

Around busy airports the terminal area is where the transition takes place between the en-route network and the approaches and departures to and from the airport. This is the most complex part of the airspace with climbing and descending traffic that needs to be separated safely. Navigation Accuracy requirements of ± 1 nmi are typically required for this phase of flight.

Approach

The approach is the final and most demanding part of the flight which delivers the aircraft to a point where the flight crew can take over visually and perform a safe landing.

Instrument Approaches used to be divided into just two categories:

- Non-Precision Approach (NPA), using the conventional navigation infrastructure such as VOR, DME or NDB to provide lateral guidance and delivering the aircraft to a point aligned to the runway centreline where the pilot could then perform a visual landing. Should the runway, or its lights, not be seen by the Missed Approach Point (MAPt) the pilot is to *go around*.
- Precision Approach (PA), based on an Instrument Landing System installed on the airport which provides both lateral and vertical guidance on a stable continuous descent path to the runway threshold. Because there is a vertical path for the pilot to follow, the minima for a PA is often much lower than for a NPA. The pilot will be head down following the instruments until decision height (or altitude) at which point he/she will look up and make a visual decision to land if the lights and/or runway is visible or, if not, fly a missed approach.

GNSS is currently only approved for use up to Category I precision approach meeting the requirements provided in the last row of Table 30.3. Work is ongoing to develop the Category II and III requirements for GNSS systems and an additional row for these applications will be added to the Table.

GNSS has enabled the introduction of a third type of approach called an RNAV approach which can include guidance in both the lateral and vertical dimensions. RNAV approaches are divided into those with only lateral guidance that are called non-precision approaches (NPA) and those with both lateral and vertical guidance that are called APV. These APV operations are neither NPA nor PA but lie in between the two existing categories. In Table 30.3 the requirements for these approaches are indicated in two rows called APV-I and APV-II. RNAV Approaches have had a significant impact, particularly on general aviation users that often fly to airfields not equipped with precision approach and landing systems.

The vertical profile of a typical approach is illustrated in Fig. 30.12. The approach is divided into four segments, Initial, Intermediate, Final and Missed approach. As already stated, in a NPA the pilot descends to a Minima Descent Altitude or Height which has been calculated by the procedure designer to ensure obstacle clearance. If at that height, the pilot has the necessary visual references to safely land the aircraft the approach is completed visually. If not, he continues to the MAPt where the go-around is to be initiated. In a precision approach or an Approach with Vertical Guidance (APV)

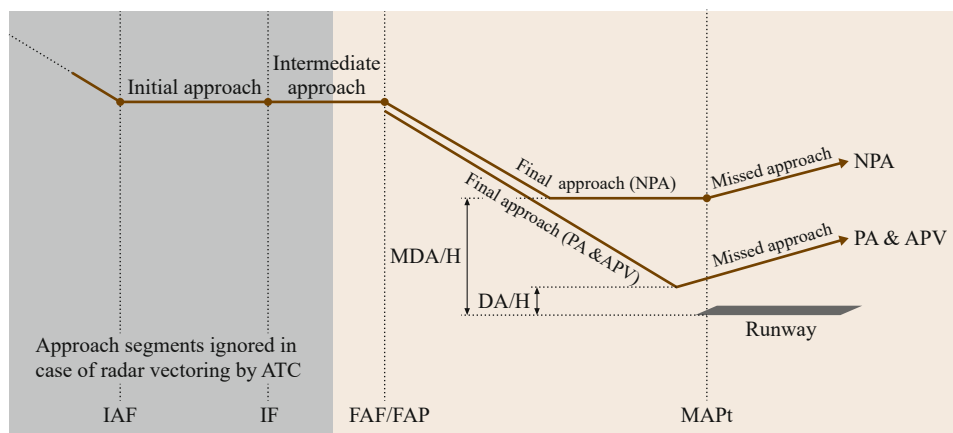


Fig. 30.12 Vertical Profile of different types of approach showing the Decision Altitude/Height for approaches with vertical guidance and, the Minimum Descent Altitude/Height for a NPA and the Missed Approach Point. The procedure is made up of a series of points from the Initial Approach Fix (IAF), the Intermediate Fix (IF), Final Approach Fix (FAF) or Final Approach Point (FAP) and the Missed Approach Point (MAPt)

the procedure designer has calculated not only a lateral path to the runway but also a vertical path. This means that the pilot has a lateral and vertical guided path providing a much higher level of accuracy on the approach and allowing the pilot to be instrument flying until he/she arrives at the Decision Height which is the point that the pilot is required to take over the landing visually or initiate a missed approach.

30.6.2 RNAV Approaches

Controlled Flight into Terrain (CFIT) occurs when a normally functioning aircraft under the complete control of the pilot is inadvertently flown into an obstacle. The pilots are generally unaware of the danger until it is too late. RNAV approaches improve safety by providing the pilot with better situational awareness, reducing the risk of CFIT and providing lower approach minima to runways not equipped with a precision approach and landing system.

All RNAV Approaches require the use of GNSS and this is reflected in the name of the procedure which is of the form: RNAV_(GNSS) RWY 27. The number at the end indicating the runway heading rounded to the nearest ten degrees which in this case would be 270°.

The different types of RNAV approach, the associated terminology and the number of acronyms used is rather complex. In the ICAO PBN concept these approaches are RNP approaches because they require on-board performance monitoring and alerting. However, by the time ICAO had defined them as RNP approaches many procedures had already been published with the title RNAV and it was impossible to go back. A transition to using the RNP terminology in the

approach chart title is now planned by ICAO and should be complete in 2023.

Different Types of RNAV Approach

There are four types of RNP approach (RNP APCH) specified in the ICAO PBN manual (Fig. 30.13). For LNAV and LNAV/VNAV procedures the lateral position and integrity is provided using GPS with RAIM. The LP and LPV procedures require the use of GPS augmented by a Satellite Based Augmentation System (SBAS).

The terms LNAV, LNAV/VNAV, LP and LPV are the labels used on the minima lines that are published on the approach chart. An example of a typical RNP approach chart is shown in Fig. 30.14. The procedure includes three minima lines for different types of aircraft capabilities. The main part of the chart is the plan view showing the waypoints, the tracks between them and some speed restrictions. Below that is the vertical profile of the procedure from the Intermediate Fix (IF) to the Final Approach Fix (FAF) and then

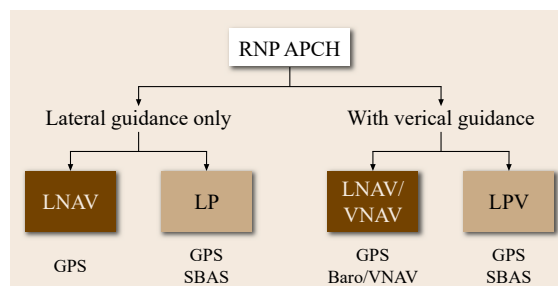


Fig. 30.13 The four RNP Approach types specified by ICAO

the final approach segment down to the missed approach point (MAPt) generally located at the runway threshold. Below that are the approach minima that depend on the aircraft equipment and operational approvals. The LPV line is for aircraft equipped with SBAS, the LNAV/VNAV line for those with Baro/VNAV vertical guidance and the LNAV for GPS lateral navigation only.

It should be noted that all such approaches are RNP approaches in the PBN terminology, because on-board monitoring and alerting is required, but the title of the chart is RNAV_(GNSS). This is a source of confusion and ICAO is proposing to evolve all chart names to use the RNP terminology in the future but this will take many years to achieve.

Performance Requirements for RNAV Approach

To support Non Precision Approaches to LNAV minima the RNP lateral accuracy in the Final Approach segment is 0.3 nmi (556 m). That means that the Total System Error (TSE) which includes both the NSE and FTE must be less than 556 m. The GNSS signal-in-space must meet the requirements in the third line of Table 30.3. The alert limit is set to 0.3 nmi and the horizontal NSE requirement is 220 m. This can be rather confusing as the 95% accuracy requirement and the Alert Limit are using the same value and yet they are not the same thing as can be seen from the definitions earlier. Setting the Alert limit to the 95% TSE requirement is actually very conservative. As the GNSS

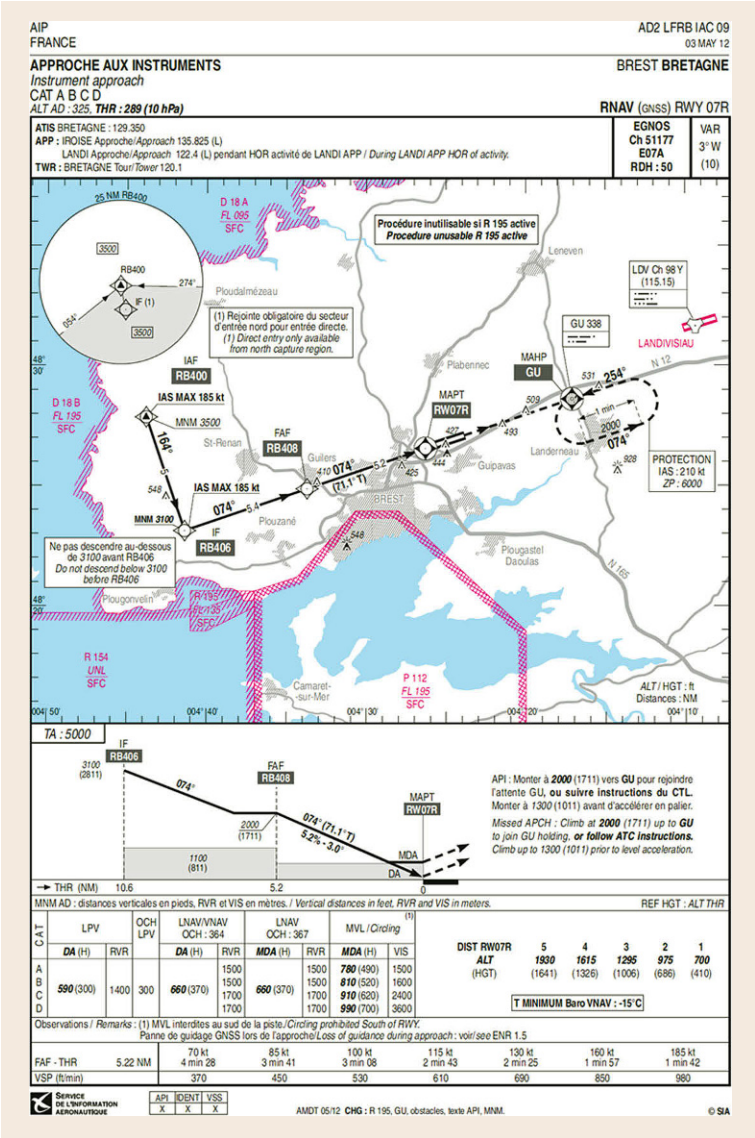


Fig. 30.14 An example of an RNAV approach chart

position estimation is so good, the majority of the error is apportioned to the FTE and an FTE of 0.25 nmi is allowed for manual flight. The LNAV/VNAV procedures use the same lateral performance requirements but add the vertical guidance based on Barometric VNAV.

For the SBAS based LPV approach there is no single RNP value in the PBN manual as the required performance is angular and therefore varies with distance from the runway threshold. The procedure is protected by obstacle clearance surfaces as illustrated in Fig. 30.15. No obstacle should penetrate these surfaces. Although the total system performance requirement is angular the performance required from the GNSS element is fixed along the final segment of the approach. The requirements are provided in Table 30.3 where it can be seen that there are three lines applicable to the LPV approach, APV-I, APV-II and Category I precision approach. The lateral accuracy and alert limit requirements for these three operations are identical. The difference is in the vertical performance and the time-to-alert. SBAS systems such as the US WAAS and the European EGNOS that have been designed to support LPV approach operations have initially targeted the APV-I performance level. The APV-II performance level is not used and will likely be removed from the ICAO standards in future updates. After gaining some experience and collecting data on the actual performance of the WAAS system and the relationship between the vertical accuracy and the alert limits the FAA assessed that a 35 m alert limit achieved the necessary performance to allow LPV operations down to 200 ft Decision Height which is equivalent to an ILS Category I procedure.

The type of RNP approach to implement at a particular aerodrome is highly dependent on the users and what their aircraft are equipped with. A significant proportion of commercial airliners from Boeing and Airbus are already equipped with Baro/VNAV capability. The first commercial airliner that offers SBAS LPV capability as an option is the Airbus A350. There are therefore very few commercial airlines with SBAS ca-

pable aircraft and so their preferred approach solution is Baro/VNAV. There is no Baro/VNAV system available for smaller general aviation type aircraft so their solution to have vertical guidance will be to equip with an SBAS receiver.

The SBAS FAS Data Block

The final approach segment of SBAS and GBAS Approach procedures are defined in a FAS data block which is stored in the on-board navigation database. The FAS data block contains all the relevant data necessary to define the final approach path and it is wrapped with a Cyclic Redundancy Check (CRC) in order to secure the integrity of the data (Fig. 30.16). The avionics receiver decodes the FAS data block and applies the CRC algorithm. If any data is changed then the CRC computation will produce a different result. If the CRC test fails then the approach is rejected.

30.6.3 RNP AR APCH

Another type of approach called RNP Approval Required (RNP AR) has been introduced which takes further advantage of GNSS performance combined with the FMS capabilities of modern aircraft and specific crew training. The performance required for RNP AR approaches are contained in a Navigation Specification in the ICAO PBN Manual [30.1]. They are not explicitly covered by Table 30.3 as they are more linked to the aircraft performance and do not impose additional requirements on the GNSS signal-in-space. The Minimum Obstacle Clearance (MOC) requirements for RNP AR approach procedures are illustrated in Fig. 30.17. The final approach requirements in the lateral dimension are equivalent to those for Non-precision approach and the vertical requirements are those for RNP Approach with Barometric VNAV.

RNP AR procedures can be used in challenging obstacle environments to allow instrument approach procedures to be implemented at locations where only visual procedures were previously possible. RNP AR is an operation where the aircraft's very tight lateral

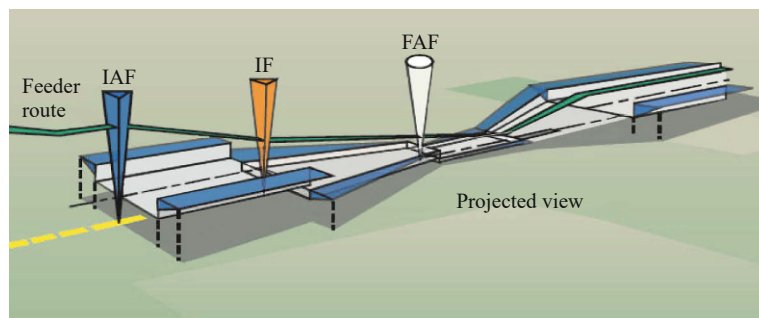


Fig. 30.15 Obstacle clearance surfaces for an approach with angular guidance on the final approach segment showing the initial approach fix (IAF), Intermediate fix (IF) and final approach fix (FAF)

Input data

Parameters	Values
Operation type	0
SBAS provider	1
Airport identifier	LFBA
Runway	29
Runway direction	0
Approach performance designator	0
Route indicator	
Reference path data selector	0
Reference path identifier	E29A
LTP/FTP latitude	441018.1420N
LTP/FTP longitude	0003603.1370E
LTP/FTP ellipsoidal height (metres)	108.8
FPAP Latitude	441039.7340N
Delta FPAP latitude (seconds)	21.5920
FPAP longitude	0003449.5365E
Delta FPAP longitude (seconds)	-73.6005
Threshold crossing height	15.00
TCH Units Selector	1
Glidepath angle (degrees)	3.30
Course width (metres)	105.00
Length offset (metres)	0
HAL (metres)	40.0
VAL (metres)	50.0

Output data

Data block	10 01 02 06 0C 1D 00 00 01 39 32 05 3C D9 F4 12 82 03 42 00 40 18 B0 A8 00 FF C0 FD 2C 81 4A 01 64 00 C8 FA 60 70 CB 84
Calculated CRC Value	6070CB84

Required additional data (not CRC wrapped)

These additional data are not required for CRC calculation, but they need to be provided to datahouses for procedure coding in ARINC 424 records.

Parameters	Values
ICAO code	LF
LTP/FTP orthometric height (metres)	61.0
FPAP orthometric height (metres)	61.0

Reset

Edit

Text report

File download

Fig. 30.16 The contents of the SBAS FAS data block

performance capabilities enable operations into very challenging terrains. Only the very high end aircraft have this level of capability and the cost to certify and operationally approve the crew is high. To qualify for RNP AR the state of ownership of the procedure must specifically authorise the Aircraft Operator (AO). The AO in turn must have demonstrated to that State’s Civil Aviation Authority (CAA) that they have the appropriate training, servicing and maintenance schedules in place to ensure that the aircraft will accurately maintain the defined path during the approach. This is very important as the designer will only protect the procedure to twice the RNP. Therefore, if the procedure was based on RNP 0.1 then there could be an obstacle at 0.2 nmi from the route.

GNSS is the only sensor that will support the level of accuracy required for these very demanding operations.

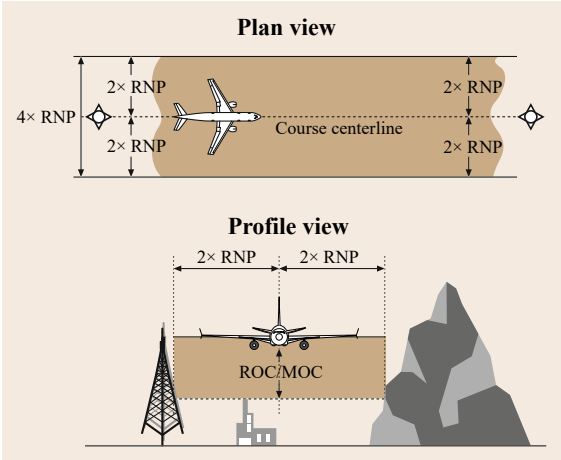


Fig. 30.17 Minimum (Required) Obstacle Clearance for RNP AR operations

30.7 Flight Planning and NOTAMs

Information about the unavailability of navigation aids is traditionally communicated in the form of Notices to Airmen (NOTAM). In preparation for a flight a pilot will consult the NOTAMs relevant to the route being flown and the destination airport to find out if there are any issues that need to be considered. Conventional navigation aids were either on or off and the impact of an outage was very clear. With GNSS the situation becomes more complex. The effect of a satellite being unavailable can vary significantly depending on the number and geometry of other satellites available. It will also vary with time. The operational impact of an outage will also depend on the avionics equipment, the mask angle being used, the satellite selection algorithms and many other factors. This makes it rather difficult to predict.

Prediction tools such as the Eurocontrol AUGUR tool [30.35] and the FAA RAIM Service Availability Prediction Tool [30.36] have been developed to forecast

GPS RAIM availability based on constellation status information that is available from the US Coast Guard. These tools are available via web interfaces for aircraft operators to verify the GPS RAIM availability at the destination airport at the expected time of arrival. In States that have published GNSS based instrument approach procedures predicted GPS RAIM outages are also provided to operators in the form of NOTAMs.

Much debate has taken place within the aviation community as to the value of GPS RAIM NOTAMs [30.37]. The predictions are not very reliable and must be tailored to the worst performing user equipment so they are inherently conservative. However, as GNSS performance varies in a way that is predictable it can also be argued that any advance warning of a possible outage would be a good thing. Work is ongoing to develop more accurate GNSS performance prediction capabilities. The debate is far from over and different solutions have been adopted in different parts of the world.

30.8 Regulation and Certification

Aviation is a highly regulated industry and all systems, equipment and procedures are governed by strictly enforced standards. As far as possible these standards are harmonised globally through ICAO. Aircraft fly all over the world and do not want to have to use different systems or follow different rules and procedures as they cross national boundaries.

30.8.1 Airworthiness Certification

A GNSS installation on an aircraft must meet the airworthiness certification standards and be approved by the appropriate regulatory authority governing the State of the aircraft manufacturer. Modifications to existing aircraft need to be approved by the regulatory authority governing the State of registry of the aircraft. In Europe airworthiness certification is performed by EASA, in the US by the FAA. As a general principle of the Chicago convention a certification in one country or region is generally accepted in another.

FAA and EASA publish Technical Standard Orders (TSOs) and European TSOs (ETSOs) respectively that are normally aligned with each other. These TSOs in turn make reference to industry standards from bodies such as RTCA and Eurocae.

The standard for basic GPS equipment is TSO C-129a [30.7] which requires RAIM fault detection as a minimum. This in turn makes reference to the Min-

imum Operational Performance Standards for Airborne Supplemental Navigation Equipment Using Global Positioning System (GPS) [30.6]. The TSO C-129a user equipment can support all operations from En-route down to NPA.

The TSO C-129 has been superseded by TSO C-196 [30.38]. The latter makes reference to DO 316 (MOPS for GPS ABAS, [30.39]), which requires the receiver to be aware that Selective Availability has been removed and employs RAIM FDE as a minimum.

Equipment with SBAS capability is addressed by the following standards:

- TSO C-145, *Airborne Navigation Sensors using GPS Augmented by SBAS* [30.40] for GNSS receivers that provide inputs to a Flight Management System,
- TSO C-146, *Stand-alone Airborne Navigation Sensors using GPS Augmented by SBAS* [30.41].

The SBAS receiver referred to in TSO C-145 [30.40] and C-146 [30.41] must meet the requirements that are specified in RTCA DO-229D [30.34] which defines different functional and operational classes of equipment depending on the specific application.

The TSOs govern the GNSS equipment itself. The integration of the equipment into the aircraft is governed by another set of standards. In the US these are

called Advisory Circulars (ACs). EASA uses the term Appropriate Means of Compliance (AMC).

In the US the Airworthiness requirements for all GNSS systems and augmentations are included in AC 20-138 [30.42].

EASA currently publishes different AMCs for each application. For example:

1. AMC 20-27: *Airworthiness Approval and Operational Criteria for RNP Approach Operations including APV Baro/VNAV operations* [30.43],
2. AMC 20-28: *Airworthiness Approval and Operational Criteria for RNAV GNSS Approach Operation to LPV Minima using SBAS* [30.44].

EASA processes are evolving and in the future the AMCs for Communications Navigation and Surveillance systems will all be combined into one single certification specification.

One negative impact of this strict certification process is that it makes any changes expensive and slow to achieve. Once an avionics system has been through the certification process it is approved for installation on-board aircraft the design is frozen. Any significant modification requires the system to be passed through the certification process again at great expense. An example of this is the slow adaptation of the avionics systems to incorporate receivers that are aware that Selective Availability (SA) has been switched off. Although SA was removed in May 2000 there are still a small number of brand new aircraft coming off the production line equipped with GPS receivers that are

not SA aware. There is still therefore, a rather large population of aircraft equipped with non-SA aware receivers. Airlines are cost driven and will only upgrade their equipment if there is a clear operational or financial benefit. Hence, aircraft are rarely retrofitted with new equipment and they tend to die with the equipment they were born with.

30.8.2 Operational Approvals

The airworthiness approval covers the aircraft and its on-board systems. In addition to having a certified aircraft an operator also requires an operational approval in order to use it for specific operations. Operational approvals are given by the National regulatory authority of the State of registry of the aircraft. In Europe the approval is based on EASA rules captured in the same AMCs as for the airworthiness certification but these rules are applied by the national regulatory authorities. In the US the operational approval is based on a series of Advisory Circulars such as AC 90-105 for RNP operations including Baro/VNAV and AC 90-107 for LPV operations.

The operational approval covers issues such as the necessary modifications to the aircraft Flight Manual, the operating guides for the on-board avionics and the flight Crew Training. It also covers flight planning procedures such as the requirement to have a non-GNSS approach procedure available as an alternate in case of a GNSS outage and the requirement to check GPS RAIM availability or NOTAMs indicating RAIM outages.

30.9 Military Aviation Applications

The military fly many different types of aircraft. Those that fly in the civilian airspace must comply, as far as possible, with the same rules and regulations as the civil aircraft they are sharing the airspace with. However, there are often exceptions.

There are also some specific aviation military applications that make use of GNSS such as the Joint Precision Approach and Landing System (JPALS) which is under development for the US military. JPALS is effectively a military version of GBAS and it comes in several versions to support different missions. There are two main categories – Shipboard Relative GPS (SRGPS) for use at sea to guide aircraft to carriers and the Local Area Differential GPS (LDGPS) for use on land. There are three different versions of LDGPS:

- Fixed base, which is very similar to GBAS.
- Tactical, which is portable for temporary installation in the field.

- Special missions – highly portable for rapid deployment in support of special operations.

JPALS will have different requirements to GBAS, in that it will need to be resistant to jamming, be difficult to detect and localise and must have the ability to operate in higher dynamic environments. This will lead to the combined use of INS/GPS integration and adaptive antenna designs as part of the solution.

The principle differences between the civil GBAS and the military JPALS are the use of:

- Dual-frequency P(Y) code GPS providing greater precision than the C/A code
- Anti-jamming technology using digital nulling and beam steering antennas
- UHF encrypted data link in place of the civil VHF datalink used in GBAS.

JPALS will also be produced in a mobile form factor suitable for rapid deployment and setup in the field. The avionics will be compatible with civil GBAS and SBAS so that military aircraft will be able to use the approach procedures available at civil airfields.

The shipboard version of JPALS, SRGPS [30.45] aims to provide a system capable of automatically landing an aircraft on a moving aircraft carrier. It uses

a relative navigation approach, the reference station being installed on the ship. In this challenging environment the system applies carrier phase tracking and the integration of inputs from inertial sensors which enables very precise positioning to be achieved relative to the desired touchdown point. Sea based JPALS is expected to reach its initial operational capability (IOC) in around 2020.

30.10 Other Aviation Applications of GNSS

30.10.1 Surveillance (ADS-B)

At the centre of an air traffic management system is the need for Air Traffic Controllers to have a real-time picture of where the aircraft are in the airspace. This picture is generally constructed by using inputs from radar stations that detect the positions of the aircraft.

Automatic Dependent Surveillance (ADS) is a surveillance technique where each aircraft automatically broadcasts its own position periodically via data-link. This is a dedicated GNSS position which is then used for ATC purposes. There are two forms of ADS, ADS Contract (ADS-C) and Broadcast (ADS-B). ADS-C by its name is a dedicated *handshake* between an aircraft and a ground station. The aircraft can only set up 5 contracts at once but the advantage is that one or both parties are made aware if transmitted data fails to be received. ADS-B is just a broadcast of data to all other stations, both air and ground, within line of sight [30.46]. ADS data can be received by ground stations and used to construct a surveillance picture for ATC which is independent from the traditional radar surveillance network. The ADS-B service also requires quality measures to be provided with the broadcast position information. Unlike conventional navigation sources the GNSS position includes the built-in integrity monitoring capabilities described earlier. It is therefore the GNSS position that is being used as the source for the broadcast position of the aircraft. There are several reasons ADS-B is being promoted not least being the possibility to reduce the number of expensive surveillance radar stations. The GNSS position is also more accurate than the current radar derived positions. In oceanic airspace today the service providers are looking at space-based ADS-B [30.47] which will use a low Earth orbit satellite constellation, such as Iridium, to receive the ADS-B transmissions and forward them on to an ATC centre.

The increased reliance on GNSS for navigation purposes and the additional use of the same information for surveillance raises the potential for common failure modes. If ATC were relying entirely on ADS-B

for surveillance which uses the same GNSS position as the aircraft is using for navigation then the impact of a GNSS outage becomes potentially much more significant. This issue needs to be considered carefully in a safety assessment.

Fortunately, in busy areas of the world there are already multiple layers of radar surveillance coverage and ADS-B is not intended to become the only means of surveillance. However, an ADS-B surveillance layer may be used to remove some of the redundancy in the current system and could lead to significant savings. In remote areas ADS-B can provide a surveillance capability where there was none previously but care must be taken to ensure that a safe contingency procedure is available to cover GNSS outages.

Multilateration (MLAT) [30.48] is another emerging surveillance technique which works by receiving the *squawk* from an aircraft's secondary surveillance radar (SSR) transponder. This squawk is received by a series of ground receivers and the location of the aircraft is computed using triangulation. MLAT position estimations rely on accurate timing and it is common for the GNSS timing pulse to provide the common time source. Loss of the GNSS signal would mean that the ground stations would slowly become unsynchronised.

30.10.2 Datalink

Air Traffic Control centres are currently implementing controller pilot data-link communications (CPDLC) which allows the exchange of text messages between pilots and controllers [30.49]. CPDLC complements the exchange of information by voice and enables a reduction in the use of the VHF communications channels. The messages exchanged in CPDLC are time stamped and many of the systems use GPS time as the reference. In order to limit the vulnerability of such systems to GPS outages, terrestrial time reference signals, such as the DCF77 long wave time signal [30.50] broadcast by the German National Physics laboratory from Mainflingen, near Frankfurt are also integrated in the system.

30.11 Future Evolution

30.11.1 GNSS Vulnerability and Alternative-PNT

One of the most significant issues around the use of GNSS in aviation is vulnerability [30.51–53]. GPS signals are low power and are broadcast on a single frequency so interference, either deliberate or unintentional, is a potential threat (Chap. 16). This has not stopped aircraft equipping with GPS receivers but it has made it difficult to show the real benefit and added value. GPS has been incorporated in addition to the existing terrestrial navigation infrastructure. There has been no cost saving due to a reduction in the number of conventional navigation aids. They have been kept as a fall back capability for the case when GPS is unavailable. The route structures being flown are not dependent on the availability of GPS and can be supported using conventional navigation aids. This is however going to change as benefits from the implementation of Performance Based Navigation start to be realised. PBN procedures already exist in a few Terminal Areas such as those around Zurich and Amsterdam Schiphol airports and will be mandatory in the majority of Europe's Terminal Airspace by January 2024. PBN procedures will make use of GPS as the primary means of navigation. In order to cope with GPS outages and maintain the level of safety an alternative positioning system needs to be available.

In the USA the term Alternative Positioning Navigation and Timing (A-PNT) is being used to describe this *back-up* to GPS [30.54]. The objective is to ensure that operations can continue safely during a GPS outage without an unacceptable increase in workload for either the pilot or the Air Traffic Controller. The FAA has launched an A-PNT programme [30.55] with the aim to determine how the alternative method of PNT can be implemented at the lowest possible cost. In Europe there have been studies and simulations to assess the impact of GNSS outages and assess the need for a back-up. These have concluded that in the near term the current network of DMEs can support a reversion capability so long as a number of operational measures are put in place. These include:

- ATC need to be informed of the geographical location and size of the GNSS outage area.
- ATC need a means to identify which aircraft are affected.
- Aircraft relying only on GPS should be blocked from entering the affected sectors.

Aircraft in the affected airspace with DME/DME RNAV systems will be able to continue to navigate

although they may have lost the ability to perform on-board performance monitoring and alerting. They will have effectively switched from RNP to RNAV. DME/DME coverage is available for all European ATS routes. The number of aircraft that are solely dependent on the use of GPS for navigation is very limited and ATC can manage these by monitoring the surveillance picture and giving radar vectors.

In the longer term it is anticipated that multi-frequency, multi-constellation GNSS will provide a higher level of robustness and the requirements on the back-up solution will diminish. However, it has not yet been demonstrated that the level of robustness will be sufficient to avoid the need for retaining a ground based alternative as a back-up.

30.11.2 Rationalisation of the Navigation Infrastructure

Ever since the arrival of GPS in aviation there have been discussions about the rationalisation of the conventional navigation aid infrastructure. The role of the conventional navigation aids is changing from being the primary means of navigation to being a support to reversion scenarios during GNSS outages. There is an expectation that this will allow a reduction to a *minimum network* that allows safety to be maintained while supporting a certain capacity of traffic in the airspace. The situation differs between the en-route and approach phases of flight.

In the en-route phase GNSS is rapidly becoming the primary navigation source and, as discussed in the previous section the favoured reversion solution for aviation is DME/DME RNAV as a sufficient number of aircraft are already equipped. The VOR and NDB will not be necessary in the future but they cannot all be removed overnight. The rationalisation of the infrastructure takes time and costs money. A broad consultation with potential users is needed and sufficient advanced warning must be given. Many published procedures are published in relation to VOR installations and these procedures would need to be replaced before the removal of the navigation facility.

Many NDB facilities are installed at aerodromes and support Non-Precision approach procedures. In order to remove the dependency on the NDB the existing procedures would need to be replaced by RNAV_(GNSS) procedures but until all users are equipped to fly the RNAV procedures then the old ones will need to be maintained. These facilities are also often used to train pilots in instrument flying as the ability to fly a non precision approach using an NDB is a nec-

essary part of the basic training syllabus which may need to be modified to change the training requirements.

In France a plan has been initiated to replace a certain number of Category I Instrument Landing Systems with SBAS LPV procedures which can achieve similar minima. Once again it has to be verified that the users of these airfields are all equipped with the SBAS capability needed to fly the LPV operations.

30.11.3 Multi-Constellation

Today the use of GNSS in civil aviation is limited to single frequency, GPS L1. Although the GPS system is broadcasting signals on another frequency (L2) this is not used in civil aviation as it is not in a protected frequency band. The introduction of the new L5 signal in GPS which is in a frequency band protected for aviation applications will allow the use of dual frequency GNSS in the future. The availability of multi-frequency, multi-constellation GNSS will bring more robustness to the navigation solution for aviation and will help to address some of the fears related to the vulnerability of the current GPS [30.56, 57].

The US GPS has been embraced by the aviation community and is providing significant operational benefits. Additional core constellations from the Russian GLONASS to the coming European Galileo

and Chinese BeiDou are going to bring further benefits [30.58, 59] but the increased number of signals and their complexity will bring some challenges, particularly in the area of standardisation.

Standards are already available in ICAO Annex 10 for the GPS and GLONASS constellations and the ICAO Navigation Systems Panel is working on the standards for Galileo and BeiDou. However, globally accepted avionics standards are currently only available for GPS user equipment. The future avionics will clearly be multi-constellation but the way to standardise the user equipment that could be using a wide variety of signals from four different constellations is a problem that has yet to be resolved.

In addition to the technical difficulties there are institutional and political issues. At an ICAO Air Navigation Conference in 2012 the Russian Federation announced that from 2017 all Russian registered aircraft would be required to carry GLONASS equipment. This raised significant concerns among the aviation community that there would be a loss of global interoperability concerning GNSS with the potential to have different forms of GNSS mandated in different regions of the world.

The community is working hard to establish a means of standardising a multi-constellation user equipment for aviation that includes all four constellations and it must be hoped that they succeed.

References

- 30.1 Performance Based Navigation (PBN) Manual, ICAO Doc. 9613 Ser., 4th edn. (ICAO, 2013)
- 30.2 P.B. Ober, D.-J. Moelker, E. Theunissen, R.C. Meijer, D. van Willigen, R. Rawlings, M. Perry: The suitability of GPS for basic area navigation, Proc. ION GPS, Kansas City (1997) pp. 1007–1018
- 30.3 K.L. Van Dyke: The world after SA: Benefits to GPS integrity, Proc. IEEE PLANS, San Diego (2000) pp. 387–394
- 30.4 K. Doucet, Y. Georgiadou: The issue of selective availability, GPS World **1**(5), 53–56 (1990)
- 30.5 K.D. McDonald: GPS in civil aviation, GPS World **2**(8), 52–59 (1991)
- 30.6 Minimum Operational Performance Standards for Airborne Supplemental Navigation Equipment Using Global Positioning System (GPS), RTCA DO-208, 07/12/1991 (RTCA, Washington DC 1991)
- 30.7 Airborne Supplemental Navigation Equipment Using the Global Positioning System (GPS), TSO-C129a (FAA, Washington DC 1996)
- 30.8 Guidance Material on Airworthiness Approval and Operational Criteria for the use of Navigation Systems in European Airspace Designed for Basic RNAV Operations, JAA Temporary Guidance Leaflet No. 2 (JAA, Hoofddorp 1996)
- 30.9 G.E. Michael: Legal issues including liability associated with the acquisition, use, and failure of GPS/GNSS, J. Navig. **52**(2), 246–251 (1999)
- 30.10 S. Malys, J. Slater: Maintenance and enhancement of the World Geodetic System 1984, Proc. ION GPS, Salt Lake City (1994) pp. 17–24
- 30.11 C. Boucher, Z. Altamimi: ITRS, PZ-90 and WGS 84: Current realizations and the related transformation parameters, J. Geod. **75**(11), 613–619 (2001)
- 30.12 ICAO: *Annex 10 to the Convention on Civil Aviation, Aeronautical Telecommunications*, Radio Navigation Aids, Vol. 1, 6th edn. (ICAO, Montreal 2006)
- 30.13 GPS Standard Positioning Service Performance Standard, 4th edn. (US Department of Defense, Washington DC 2008)
- 30.14 V. Iatsouk: Development of standards for aeronautical satellite navigation system, Acta Astronaut. **54**(11), 961–963 (2004)
- 30.15 W.Y. Ochieng, K. Sauer, D. Walsh, G. Brodin, S. Griffin, M. Denney: GPS integrity and potential impact on aviation safety, J. Navig. **56**(1), 51–65 (2003)
- 30.16 D. Lawrence, D. Bunce, N.G. Mathur, C.E. Sigler: Wide Area Augmentation System (WAAS), Program Status, ION GNSS, Fort Worth (2007) pp. 892–899

- 30.17 P. Feuillet: EGNOS program status, ION GNSS, Nashville (2012) pp. 1017–1033
- 30.18 T. Sakai, H. Tashiro: MSAS status, ION GNSS, Nashville (2013) pp. 2343–2360
- 30.19 K.N.S. Rao: GAGAN – The Indian satellite based augmentation system, Indian J. Radio Space Phys. **36**(4), 293 (2007)
- 30.20 S. Karutin: SDCM program status, ION GNSS, Nashville (2012) pp. 1034–1044
- 30.21 A.A. Herndon, M. Cramer, K. Sprong: Analysis of advanced flight management systems (FMS), flight management computer (FMC) field observations trials, radius-to-fix path terminators, Proc. IEEE/AIAA 27th Digit. Avion. Syst. Conf., St. Paul (2008) pp. 2.A.5–1–2.A.5–15
- 30.22 ARINC 424–20, Navigation System Database Standard (Aeronautical Radio, Annapolis 2011)
- 30.23 RTCA DO 200A/Eurocae ED76: Standards for Processing Aeronautical Data (1998)
- 30.24 RTCA DO 201A/Eurocae ED77: Standards for Aeronautical Information (2000)
- 30.25 B. Haltli, P. Ewing, H. Williams: Global navigation satellite system (GNSS) and area navigation (RNAV) benefiting general aviation, Proc. 24th Digit. Avion. Syst. Conf., Crystal City (2005) pp. 13.A.5–1–13.A.5–8
- 30.26 Roadmap for Performance Based Navigation, Evolution for Area Navigation (RNAV) and Required Navigation Performance (RNP) Capabilities 2006–2025, Version 2.0 (FAA, Washington DC 2006)
- 30.27 European Airspace Concept Handbook for PBN Implementation, 3rd edn. (Eurocontrol, Brussels 2013)
- 30.28 K. Kovach: Continuity: The hardest GNSS requirement of all, Proc. ION GPS, Nashville (1998) pp. 2003–2020
- 30.29 I. Mallett, K. Van Dyke: GPS availability for aviation applications: How good does it need to be?, Proc. ION GPS, Salt Lake City (2000) pp. 705–712
- 30.30 R.G. Brown: A baseline GPS RAIM scheme and a note on the equivalence of three RAIM methods, Navigation **39**(3), 301–316 (1992)
- 30.31 J.P. Fernow, Y.C. Lee: Analysis supporting FAA decisions made during the development of TSO C-129, Proc. ION AM 1994, Colorado Springs (1994) pp. 219–228
- 30.32 P.B. Ober: RAIM Performance: How Algorithms Differ, ION GPS 1998, Nashville 15–18 Sep. 1998 (ION, Virginia 1998) pp. 2021–2030
- 30.33 A. Martineau, Ch. Macabiau, M. Mabilieu: GNSS RAIM assumptions for vertically guided approaches, ION GNSS 2009, Savannah Sep. 2009 (ION, Virginia 2009) pp. 2791–2803
- 30.34 Minimum Operational Performance Standards for Global Positioning System/Wide Area Augmentation System Airborne Equipment, RTCA DO229D, 13/12/2006 (RTCA, Washington DC 2006)
- 30.35 D.A.G. Harriman, J. Wilde, P.B. Ober: EUROCONTROL's predictive RAIM tool for en-route aircraft navigation, IEEE Aerosp. Conf. 1999, Snowmass at Aspen 6–13 Mar. 1999 (IEEE, New York 1999) pp. 385–393
- 30.36 ADS-B Service Availability Prediction Tool Receiver Autonomous Integrity Monitoring User Guide, v2.0, 30 Apr. 2014 (FAA, Washington DC 2014)
- 30.37 Massimini, V. McNeil, G. Scales, W.: *Proposed Concept of Operation for a GNSS NOTAM and Aeronautical Information System* (The MITRE Corporation, Bedford 2008)
- 30.38 Airborne Supplemental Navigation Sensors for Global Positioning System Equipment Using Aircraft Based Augmentation, TSO-C196 (FAA, Washington DC 2009)
- 30.39 Minimum Operational Performance Standards for Global Positioning System/Aircraft Base Augmentation System, RTCA DO-316 (RTCA, Washington DC 2009)
- 30.40 Airborne Navigation Sensors Using the Global Positioning System Augmented by the Satellite Based Augmentation System, TSO-C145c (FAA, Washington DC 2008)
- 30.41 Stand-alone Airborne Equipment Using the Global Positioning System Augmented by the Satellite Based Augmentation System, TSO-C146c (FAA, Washington DC 2008)
- 30.42 Airworthiness Approval of Positioning and Navigation systems, FAA Advisory Circular (AC), 20-138D, 28/03/2014 (FAA, Washington DC 2014)
- 30.43 European Aviation Safety Agency: Airworthiness Approval and Operational Criteria for RNP Approach (RNP APCH) Operations Including APV Baro/VNAV Operations, AMC 20-27 (EASA, Cologne 2009)
- 30.44 European Aviation Safety Agency: Airworthiness Approval and Operational Criteria for RNAV GNSS Approach Operation to LPV Minima Using SBAS, AMC 20-28 (EASA, Cologne 2012)
- 30.45 K.L. Gold, A.K. Brown: A hybrid integrity solution for precision landing and guidance, IEEE PLANS 2004 (IEEE, New York 2004) pp. 165–174
- 30.46 C. Rekkas, M. Rees: Towards ADS-B implementation in Europe, Proc. Tyrrhenian Int. Workshop Digit. Commun.-Enhanc. Surveill. Aircr. Veh. (TI-WDC/ESAV), Capri (IEEE, New York 2008)
- 30.47 T. Delovski, K. Werner, T. Rawlik, J. Behrens, J. Bredemeyer, R. Wendel: ADS-B over satellite – The world's first ADS-B receiver in space, 45 Small Satell. Syst. Serv. Symp. 2014, ESA, Noordwijk (2014)
- 30.48 N. Xu, R. Cassell, C. Evers, S. Hauswald, W. Langhans: Performance assessment of Multilateration Systems – A solution to nextgen surveillance, Proc. Integr. Commun. Navig. Surveill. Conf. (ICNS'10), Herndon (IEEE, New York 2010), pp. D2-1–D2-8
- 30.49 C. Collings, J. Harwood: Data link messaging standards for NextGen data communications, Proc. Integr. Commun. Navig. Surveill. Conf. (ICNS'09), Arlington (IEEE, 2009)
- 30.50 Time and Standard Frequency Station DCF77 (Germany), <http://www.eecis.udel.edu/~mills/ntp/DCF77.html>
- 30.51 J.V. Carroll: Vulnerability assessment of the US transportation infrastructure that relies on the global positioning system, J. Navig. **56**(2), 185–193 (2003)
- 30.52 D. Last: GPS forensics, crime, and jamming, GPS World **20**(10), 8–12 (2009)
- 30.53 C. Dixon, S. Smith, A. Hart, R. Keast, S. Lithgow, A. Grant, J. Šafář, G. Shaw, C. Hill, S. Hill, C. Betty:

- Specification and testing of GNSS vulnerabilities, Proc. ENC-GNSS 2013, Vienna (ENC, Vienna 2013) pp. 1–12
- 30.54 E. Kim: Investigation of APNT optimized DME/DME network using current state-of-the-art DMEs: Ground station network, accuracy, and capacity, IEEE/ION PLANS 2012, Myrtle Beach (IEEE, New York 2012) pp. 146–157
- 30.55 Concept of Operations for NextGen Alternative Positioning, Navigation and Timing (APNT) (FAA, Washington DC 2012)
- 30.56 C.J. Hegarty, E. Chatre: Evolution of the Global Navigation Satellite System (GNSS), Proc. IEEE **96**(12), 1902–1917 (2008)
- 30.57 J. Blanch, T. Walter, P. Enge: Satellite navigation for aviation in 2025, Proc. IEEE **100**, 1821–1830 (2012)
- 30.58 F. Salabert: Operational benefits of multi-constellation dual frequency GNSS for aviation, Coordinates **11**(3), 43–45 (2015)
- 30.59 B. Bonet, I. Alcantarilla, D. Flament, C. Rodriguez, N. Zarraoa: The Benefits of Multi-constellation GNSS: Reaching up Even to Single Constellation GNSS Users, ION GNSS 2009, 22–25 Savannah (ION, Virginia 2009) pp. 1268–1280

31. Ground Based Augmentation Systems

Sam Pullen

This chapter explains the fundamentals of ground-based augmentation systems (GBASs). GBASs are fielded at airports to support civil aviation operations down to and including precision approach and landing. This chapter describes how GBAS generates differential corrections for Global Positioning System (GPS) pseudorange (L1 C/A-code) signals based on measurements taken at known (reference) locations, how reference measurements are monitored to protect against GPS and GBAS faults or anomalies, and what information is broadcast to users to support enhanced accuracy and integrity (or safety). The application of GBAS to civil aviation precision approach and landing is explained along with the key considerations in fielding GBAS reference equipment at airports. Augmentation systems that transmit additional global navigation satellite system (GNSS)-like ranging signals to users are also briefly introduced.

31.1	Components	906
31.2	An Overview of Local Area Approaches	907
31.2.1	Pseudorange Corrections	907
31.2.2	Carrier-Phase Corrections	908
31.2.3	Reference Station Distribution	908
31.2.4	Broadcast Techniques	908
31.3	Ground-Based Augmentation Systems	909
31.3.1	Overview and Requirements	909
31.3.2	Generation of Differential Corrections	910
31.3.3	Fault Monitoring	911
31.3.4	User Processing and Integrity Verification	917
31.3.5	Additional Threats: RF Interference and Ionosphere	921
31.3.6	Equipment and Siting Considerations	925
31.3.7	Typical GBAS Errors and Protection Levels	926
31.3.8	Existing GBAS Ground Systems and Airborne Equipment	928
31.4	Augmentation via Ranging Signals Pseudolites	928
31.4.1	Origins and Use in Local-Area DGNSS	928
31.4.2	New-Generation Pseudolite Systems for Commercial Applications	929
31.5	Outlook	930
	References	930

Terrestrial augmentation of GNSSs can take the form of differential correction broadcasts and/or the broadcast of GNSS-like ranging signals from ground locations by so-called *pseudolites*. The original motivation for adding pseudolites to GNSSs was to compensate for the limited number of GPS satellites in the sky. For example, over-the-air tests of GPS transmitters at the Yuma test ground in the late 1970s were used to verify the performance of the GPS signal concept before sufficient satellites were in orbit to obtain position fixes solely from space [31.1]. While GPS today almost always provides four or more visible satellites above 10° of elevation, users with restricted sky visibility may still benefit from local ground-based transmitters [31.2].

Pseudolites have also been used to improve ranging geometry in such a way that carrier-phase integer ambiguities can be resolved once a user passes nearby [31.3].

The original motivation for broadcasting differential corrections to GNSS ranging signals came from the deliberate degradation of the timing accuracy of the GPS L1 C/A-code signals, which was known as selective availability, or S/A [31.4]. Error on the order of 10–20m was artificially induced on the clock component of each GPS satellite (when receiving C/A-code) in order to prevent civil users from obtaining greater accuracy than was originally intended by the system designers. The error induced by S/A on a given satellite was the same for all C/A-code users tracking that

satellite. It was soon realized that receivers located at fixed and known (precisely surveyed) sites could estimate the error in each satellite ranging measurement and, by broadcasting this information, allow nearby users to remove the impact of S/A from their measurements [31.5]. This relatively simple means of defeating S/A was one reason that the US decided to deactivate S/A in May of 2000 [31.6].

Well before S/A was deactivated, the other advantages of what became known as *differential GPS* had become apparent. Once S/A was removed, other natural error sources such as satellite clock and ephemeris errors and ionospheric and tropospheric delays became significant contributors to GPS navigation accuracy. These are all highly correlated over short to medium distances; thus the application of differential corrections removes most of their effects. In addition, differential GPS was a key component of the task of verifying the integrity, or safety of use, of GPS signals using individual ground stations or networks of stations. Ground

systems that combine differential corrections with real-time integrity information became known as *GNSS augmentation systems*.

This chapter will focus on augmentation systems that are based on a single ground station site and dissemination of corrections and integrity information via radio link. This approach is generally denoted as local area differential GNSS, or LADGNSS. The adaptation of this technology for civil aircraft precision approaches and landings, known internationally as ground-based augmentation systems (GBAS) and as local area augmentation systems (LAAS) in the US, will be discussed in detail, as it covers almost all of the key considerations that apply to single-reference-site (SRS) augmentation systems. In particular, LAAS and GBAS have been approved to support precision approaches (with assured lateral and vertical guidance) down to Category I criteria – 200 ft above the surface – and extensions of this technology are expected to support approaches all the way to the ground (Category IIIB).

31.1 Components

Figure 31.1 shows the primary components of a GBAS reference station sited at an airport [31.7]. GBAS includes multiple (four or more) ground reference receivers and antennas for redundancy and integrity monitoring. Each receiver antenna is located in a carefully presurveyed site and is tested in advance to confirm that the multipath characteristics of the site selected are bounded by predetermined models that are used in determining the standard deviations (or *sigmas*) of the errors in the pseudorange corrections that are broadcast to aircraft. These antenna sites are separated by enough distance for their multipath errors (from ground reflections and distant obstacles) to be approximately statistically independent, allowing averaging of these errors under nominal conditions and making it easier to distinguish faulty measurements in a single reference receiver. For GBAS, it is important to minimize multipath errors at these receiver antennas because these errors will not be correlated with user ranging errors and thus will not be mitigated by differential corrections. Therefore, specialized ground antenna designs are used, and these will be described further in Sect. 31.3.

In GBAS, a master control processor typically located near or next to the reference receivers collects the receiver measurements and determines the pseudorange corrections and correction rates to broadcast for each satellite approved for use. This includes determining, using a series of monitor algorithms and

exclusion logic, which satellite and ground measurements are safe for the intended purpose, as will be discussed further in Sect. 31.3. This information is updated twice per second and includes (at lower update rates) information on station characteristics as well as the ideal approach geometries to be followed for all allowed precision approaches at this airport. It is packaged into several different message types, transferred by wire or optical fiber to a very high frequency (VHF) data broadcast (VDB) transmitter and antenna somewhere on airport property, and broadcast to users using the instrument landing system (ILS) localizer band from 108–118 MHz [31.8]. Siting of the VDB antenna at each airport is key to maximizing the airport surface area where the signal can be received while also ensuring acceptable signal levels throughout the precision approach coverage region in the air.

Because of the very high demands on GBAS for integrity and continuity (safety) and availability (the percentage of time operations are possible), multiple degrees of redundant equipment are essential. The same is not true of many less-demanding uses of LADGNSS. Many short-duration applications (such as day-long survey activities) can be adequately supported by a single reference receiver and antenna. In these cases, the equipment can be checked out before and after each application, and the penalty for a failed application is limited to repeating it after the equipment is

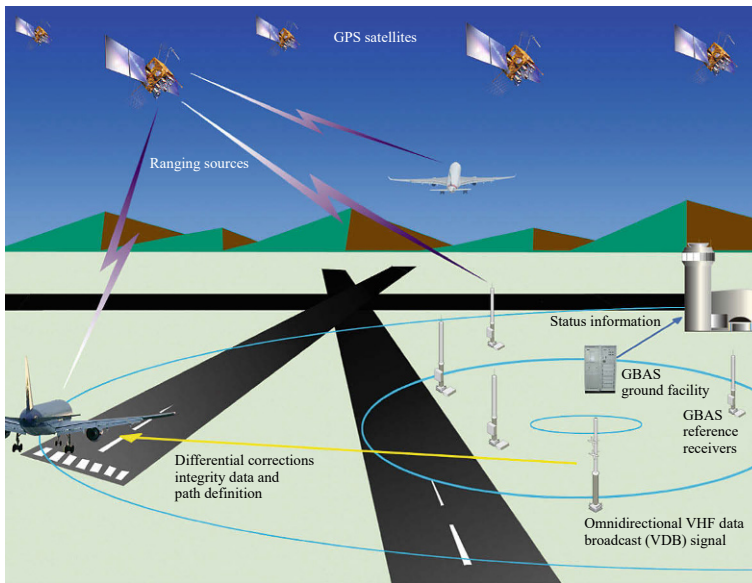


Fig. 31.1 GBAS ground station components—advanced example of terrestrial local area differential GNSS (DGNSS) systems (courtesy of the FAA satellite navigation team)

fixed. Permanent, fixed local area differential GNSS (LADGNSS) reference stations or networks will likely possess at least some redundancy to minimize mainte-

nance demands, but practically all LADGNSS applications other than GBAS are likely to need a subset of the equipment needed for GBAS.

31.2 An Overview of Local Area Approaches

While the hardware components for most LADGNSS systems are similar, significant differences exist between the algorithms used to generate differential corrections and the specific measurements that are corrected. This section will describe a few of these variations and refer to other chapters of this volume for more details.

31.2.1 Pseudorange Corrections

Almost all LADGNSS systems broadcast either corrections to pseudorange measurements or pseudorange measurements themselves so that users can remove common-mode errors by applying the broadcast information to their own measurements. In the most common case, where pseudorange corrections are sent, these corrections express the ground's best estimate of the pseudorange error on each visible satellite as measured by nearby users based on knowledge of the true positions of the reference receiver antenna(s) and the satellite locations included in the navigation data. Users then simply subtract this correction from their own pseudorange measurements to arrive at improved ones that are used in positioning calculations (Chap. 26). Note that a common procedure is required between ref-

erence and user systems to ensure that common terms are applied or not applied at both ends. GBAS, for example, applies satellite clock corrections (based on navigation data) to its reference measurements. These corrections will be the same for reference and user receivers. To maintain proper error cancellation, GBAS users must also apply satellite clock corrections but not relativistic corrections [31.9].

When a *corrections* approach similar to GBAS is used, the broadcast corrections should be close to zero and should approximate the actual error on the pseudoranges measured by users due to satellite (clock and ephemeris) and atmospheric (ionosphere and troposphere) causes. The small magnitude of this number makes it easy to broadcast with a small dynamic range; i.e., a small number of bits in a digital message format (one value for each usable satellite). Because standalone user errors are often dominated by ionospheric delay, pseudorange corrections under nominal conditions are typically their largest for low-elevation satellites when the ionosphere is particularly active (Chap. 39). The maximum magnitude of GBAS pseudorange corrections is ± 327.67 m [31.8], but anything larger than ± 125 to 150 m suggests a potential satellite anomaly, as this exceeds the range

of ionospheric delays even under extreme conditions.

Due to the time delay that exists between the measurements used to generate differential corrections on the ground and the application of the resulting corrections to user measurements, pseudorange rate corrections or other means to interpolate over the resulting latency are needed. GBAS provides a time tag for each pseudorange correction and broadcasts pseudorange rate corrections based upon the difference between the last two pseudorange corrections divided by the time interval. Users apply this pseudorange correction rate to linearly extrapolate forward from the correction time tag to the time of their measurements as long as this latency is within a time to alert that is protected by the ground station. Satellites suffering from *clock* failures can have range measurements that vary wildly and nonlinearly over a few seconds to minutes; thus these faults need to be detected quickly to protect the linear extrapolation performed by GBAS users.

31.2.2 Carrier-Phase Corrections

In addition to broadcasting pseudorange corrections, some LADGNSS systems also transmit raw carrier-phase measurements or corrections to them in support of users applying what is known as carrier-phase differential GNSS (CDGNSS) or real time kinematic (RTK) GNSS processing. CDGNSS users include those with centimeter- or even millimeter-level positioning accuracy requirements, such as surveyors and scientific users.

Carrier-phase measurements contain unknown integer ambiguities (i.e., integer numbers of sine-wave cycles) and thus do not provide unambiguous range estimates. Solving for these ambiguities is a key component of most algorithms applied by CDGNSS and RTK users, and it is this step that provides centimeter-level or better range and position accuracy (Chap. 23 and [31.3]). Some CDGPS techniques do not depend upon always resolving integer ambiguities (*fixed* solutions) but instead upon refining *floating* estimates of these ambiguities over time; thereby improving upon the performance that can be obtained from pseudorange-based LADGNSS alone. For examples of these techniques used for aviation, see [31.10, 11]. Existing GBAS systems do not broadcast carrier-phase corrections, but because of these potential benefits, Message type 6 has been reserved to support future carrier-phase corrections in the GBAS interface control document (ICD) [31.8].

31.2.3 Reference Station Distribution

Most LADGNSS systems derive and broadcast corrections from essentially a single location. This includes GBAS and other systems with reference receiver and antenna siting redundancy on a local scale (i.e., within several kilometers), because antennas this close together are observing almost identical GNSS satellite and atmospheric behavior under normal conditions. The alternative technique of distributing reference stations widely (hundreds of kilometers) apart and developing corrections that can be used over entire countries or continents is normally supported by corrections broadcast from space (so-called *space-based augmentation systems*, or SBAS) and is described in Chap. 12. In between these extremes are terrestrial LADGNSS systems, which also use networks of reference stations instead of single sites. These networks are defined as *terrestrial* instead of *space-based* because they broadcast corrections using ground-based means (Sect. 31.2.4). Station separations in terrestrial networks tend to be shorter than those in space-based networks, but this is not always the case.

The primary motivation for LADGNSS reference station networks is to provide service to a relatively large area without the need to place a large number of individual reference stations, as demonstrated for CDGNSS users in [31.12]. One system of this type conceived for aviation use in Australia was the ground-based regional augmentation system, or GRAS, described in [31.13]. It married the network concept of multiple widely-spread reference stations transmitting corrections to a central processor (as in SBAS) with the GBAS method of broadcasting corrections via terrestrial VHF radio on the ILS localizer band. The GBAS message format from [31.8] was extended with an additional message type (*type 101*) to accommodate GRAS, while the single-transmitter approach of GBAS was replaced with a network of VHF transmitter stations that already existed in Australia [31.14]. This *hybrid* approach, while not actually put into operation, illustrates the many possible variations of LADGNSS and wide-area GNSS technology that can be applied to support particular service providers and classes of users.

31.2.4 Broadcast Techniques

The most common means of communicating LADGNSS corrections to users is by radio frequency (RF) transmissions using separate transmitters and receivers. The use of the VHF ILS localizer band from 108–118 MHz for this purpose in GBAS allows the reuse of ILS localizer antennas that already

exist on almost all aircraft that will be equipped with GBAS [31.7, 8]. An RF technique commonly used in marine LADGNSS systems is digitally encoding messages on existing direction-finding radiobeacons transmitting at around 300 kHz, as is done by the US Coast Guard NDGPS [31.5, 15]. Ultra high frequency (UHF) radiomodems are commonly used in relatively low-cost commercial and experimental systems. An example of a radiomodem transmitter used in unmanned aerial vehicle (UAV) LADGNSS flight tests conducted by KAIST in Daejeon, Korea is the IP-921 made by Microhard Systems, Inc [31.16]. It transmits at a frequency range of 902–928 MHz with a maximum data rate of 1.1 Mbps out to a line-of-sight range of 100 km or more at a lower data rate of 345 kbps, which

should be completely adequate for single-reference-site LADGNSS [31.17].

Now that Internet connectivity is widespread, disseminating corrections via the Internet is also possible for some applications that do not require the highest levels of timeliness or reliability. The corrections of some existing GBAS and SBAS systems are available on the Internet after some delay. For example, the Federal Aviation Administration (FAA) William J. Hughes Technical Center, which oversees the fielding and operations of GBAS (LAAS) sites in the Conterminous United States (CONUS), includes near-real-time and archived corrections and other broadcast information for several CONUS GBAS sites on a website for analysis [31.18].

31.3 Ground-Based Augmentation Systems

31.3.1 Overview and Requirements

As described previously, GBAS is a subset of LADGNSS that supports civil aviation precision approach and landing operations at airports where a GBAS ground station is sited. The key components of a GBAS ground station are shown in Fig. 31.1 and are described briefly in Sect. 31.1. This section will describe in more detail the requirements that GBAS must meet and how GBAS satisfies these requirements through a combination of high-accuracy differential corrections, monitoring of possible faults and anomalies to protect integrity, and optimization to maximize availability and continuity in the presence of these anomalies.

Figure 31.2 expands upon Fig. 31.1 to show how the GBAS ground system components are typically sited relative to the territory of a given airport and how they support precision approaches to all runway ends at that airport. In existing sites, the GBAS reference receivers

and antennas are placed relatively close together but just far enough apart to keep multipath errors at each antenna uncorrelated, as discussed in Sect. 31.1. The central processing unit that determines the information to be broadcast is also nearby, but the VHF transmitter antenna may be some distance away to cover as much of the airport and its surrounding approaches as possible. Note that most airports have multiple runways and can at least theoretically support precision approaches in two directions (two runway ends) for each runway. Approaching aircraft are supported by GBAS to at least the Category I approach minima at all runway ends for which GBAS precision guidance is allowed (and path-guidance information is broadcast). Aircraft using GBAS derive Cartesian (e.g., east-north-up) positions from their corrected GNSS measurements and convert these to angular *ILS-lookalike* measurements to be compatible with existing avionics built for the angular measurements (offsets from desired vertical glidepath and lateral localizer) provided by the instrument landing system (ILS) [31.7, 9].

The development of GBAS from LADGNSS over the last 25 years has been spearheaded by aviation service providers such as the FAA in the US, companies that provide equipment for civil aviation such as Honeywell, Raytheon, Rockwell Collins, and Thales Group, and universities and other research groups. The requirements for GBAS have been developed from earlier systems such as ILS by Radio Technical Commission for Aeronautics (RTCA) in the US, European Organisation for Civil Aviation Equipment (EUROCAE) in Europe, and the International Civil Aviation Organization (ICAO), which devises *Standards and Recommended Practices* (SARPS) for various aircraft systems [31.7–9].

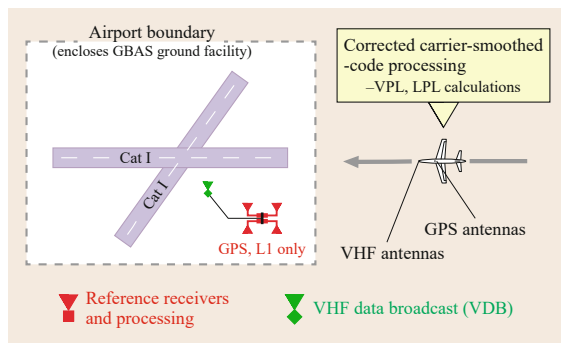


Fig. 31.2 GBAS ground station and airborne operations at a typical airport (Category I precision approach operations)

In practice, RTCA and EUROCAE standards are *harmonized* at ICAO so that GBAS ground and airborne designers work to a single, unified set of requirements to the greatest extent possible.

Table 31.1 summarizes the requirements for different modes of aviation precision approach under limited visibility as defined for GBAS [31.19]. The term *GSL* in the left-hand column stands for *GBAS Service Level* and refers to different *categories* of precision approach, which correspond to different levels of pilot visibility in horizontal (*RVR*, runway visual range) and vertical (*ceiling*). Current GBAS systems primarily support *GSL C*, which corresponds to *Category I* precision approaches down to a 200 ft minimum decision height above the runway threshold. *GSL A* and *B* support less-demanding variations of *Category I* precision approach that can be supported, with slightly different values, by *SBAS* (Chap. 12 and [31.20]). *GSL D*, *E*, and *F* refer to versions of *Category II* and *III* precision approaches that support lower decision heights. GBAS ground systems and airborne equipment that support *GSL D* and allow for *Category III* operations (including zero-visibility landings) are now under development [31.21, 22].

The columns in Table 31.1 express requirements in terms of *accuracy*, *integrity*, and *continuity*. The *accuracy* requirement is expressed as a 95th-percentile bound on navigation system error (NSE), or the difference between true (unknown) position and that output by a GBAS-qualified airborne receiver. *Continuity* is expressed as an upper bound on the probability (per 15 or 30 s exposure interval) that an operation in progress is unexpectedly aborted due to a GBAS system interruption, which can take one of several forms. *Integrity* generally refers to the trust that can be placed in the outputs of a system and the ability of a system to alert when its outputs cannot be trusted (i.e., are unsafe to use). For GBAS, this is expressed by three related requirements. The first is an upper bound on the probability (per exposure interval – typically the length of a precision approach, or about 150 to 200 s) that integrity is lost, meaning that unsafe or *misleading* information is output. The second is a *time to alert* within which a warning from the system needs to be output to avoid an unsafe condition. In other words, if the output of misleading information lasts for less than the time to alert before the system alerts that the information should not be used. This condition is not unsafe and does not count against the integrity loss probability. The third requirement, the *alert limit*, expresses the maximum error in both vertical and lateral dimensions that is regarded as *safe* for a given operation [31.23].

Availability requirements are not shown in Table 31.1 because they vary widely by airport. The

most common definition of availability is the percentage of time that the accuracy, integrity, and continuity requirements are met simultaneously, allowing GBAS-supported operations to take place. When one or more of these requirements are violated, this will be known to users and operators so that operations can be avoided. If operations begin and then must stop suddenly due to an unexpected change of system state (e.g., the failure of a GNSS satellite needed for acceptable performance, which is known as a *critical satellite*), *continuity* is lost. This means that implementing the integrity monitoring needed to meet the integrity requirements unavoidably increases the probability of continuity loss. Since integrity and continuity requirements thus compete directly with each other, the design of integrity monitoring algorithms and measurement-exclusion logic is a much greater challenge than it appears at first glance.

31.3.2 Generation of Differential Corrections

As described in Sect. 31.3.1, GBAS generates pseudorange differential corrections and correction rates for each satellite from redundant reference receiver measurements. This section provides the specified equations used to compute these values to clarify the procedure [31.24].

The first step in processing pseudorange and carrier-phase measurements from each reference receiver is to perform what is known as carrier smoothing, which takes advantage of the greater precision of carrier-phase measurements to attenuate noise and multipath errors in the pseudorange measurements (Sects. 22.3.1 and 20.4). To maximize commonality in smoothing between ground and airborne equipment, a specific algorithm is mandated for the ground system, and the airborne is given a very small tolerance for variation given nominal input conditions [31.9]. This typically requires that the same smoothing time of 100 s is used in both ground and airborne systems. The ground smoothing algorithm (at least for FAA LAAS systems) is given by [31.24]

$$\begin{aligned} \text{PR}_s(k) = & \left(\frac{1}{N} \right) \text{PR}_r(k) \\ & + \left(\frac{N-1}{N} \right) [\text{PR}_s(k-1) + \phi(k) - \phi(k-1)], \end{aligned} \quad (31.1)$$

where PR_r is the raw (measured) pseudorange for a given GNSS satellite at a given reference receiver, PR_s is the output smoothed pseudorange, f is the carrier-phase (accumulated Doppler) measurement, k represents the current time epoch, and $N = S/T$ (nom-

Table 31.1 GBAS precision approach requirements summary

GSL	Accuracy		Integrity				Continuity
	95% Lat. (NSE) (m)	95% Vert. NSE (m)	Pr (loss of integrity) (s)	Time to alert (s)	LAL (m)	VAL (m)	Pr (loss of continuity) (s)
A	16	20	$2 \cdot 10^{-7}/150$	6	40	50	$8 \cdot 10^{-6}/15$
B	16	8	$2 \cdot 10^{-7}/150$	6	40	20	$8 \cdot 10^{-6}/15$
C	16	4	$2 \cdot 10^{-7}/150$	6	40	10	$8 \cdot 10^{-6}/15$
D	5	2.9	$10^{-9}/15$ (vert.); 30 (lat.)	2	17	10	$8 \cdot 10^{-6}/15$
E	5	2.9	$10^{-9}/15$ (vert.); 30 (lat.)	2	17	10	$4 \cdot 10^{-6}/15$
F	5	2.9	$10^{-9}/15$ (vert.); 30 (lat.)	2	17	10	$2 \cdot 10^{-6}/15$ (vert.); 30 (lat.)

NSE – navigation system error

inally 200 at filter steady-state) represents the number of samples used in the smoothing filter, where T is the measurement sampling interval (nominally 0.5 s), and $S = 100$ s.

The smoothed pseudoranges from (31.1) are the basis of the pseudorange corrections (PR_{sc}) computed for each reference receiver m and each satellite n [31.24]

$$PR_{sc}(n, m) = R(n, m) - PR_s(n, m) - t_{sv_gps}(n), \quad (31.2)$$

where R is the predicted range based on the surveyed antenna position and the satellite position provided by the broadcast ephemeris data, and t_{sv_gps} is the clock correction for satellite n (for L1 C/A-code) as computed from the GPS navigation data, including relativistic terms. Note that neither ionospheric nor tropospheric corrections are applied. As mentioned in Sect. 31.2.1, corrections that would be the same (or very nearly the same) at both ground and aircraft must either be applied in both places or neither place to avoid introducing common errors.

The next step for GBAS is to perform a *clock adjustment* to the corrections generated by each reference receiver. This is done to remove the large common bias created by reference receiver clock errors from the broadcast corrections and to allow straightforward comparison or corrections across reference receivers (all variables apply for a specific epoch k) [31.24]

$$PR_{sca}(n, m) = PR_{sc}(n, m) - \frac{1}{N_c} \sum_{n \in S_c} PR_{sc}(n, m), \quad (31.3)$$

where PR_{sca} is the resulting smoothed and clock-adjusted correction for receiver m and satellite n , S_c is the *common set* of satellites, meaning the set of satellites tracked by all healthy reference receivers, and N_c

is the number of satellites in set S_c . For each receiver m , (31.3) removes the average bias in the nonadjusted pseudorange correction across the N_c satellites in the common set. This works well because the receiver clock bias on each receiver is the same for all satellites tracked by that receiver.

Under nominal conditions when no receivers are removed due to detected anomalies, the broadcast pseudorange correction (PR_{corr} or PRC) for each satellite n is simply the average over the $M(n)$ smoothed and clock-adjusted corrections from each receiver tracking that satellite [31.24]

$$PR_{corr} = \frac{1}{M(n)} \sum_{m \in S_n} PR_{sca}(n, m). \quad (31.4)$$

The broadcast range rate correction (RR_{corr} or RRC) for satellite n is derived by differencing the current and immediately previous pseudorange correction values for that satellite divided by the time gap T between them (nominally 0.5 s).

Figure 31.3 shows a block-diagram representation of the functions carried out by the GBAS ground system. It corresponds specifically to the integrity monitor testbed (IMT) developed at Stanford University as a ground system prototype and algorithm development tool [31.25], but the same (or very similar) functions are carried out by all GBAS ground systems. The functions shown in Fig. 31.3 that have been described in this section include **SISRAD** (signal-in-space receive and decode), smooth (31.1), correction ((31.2) and (31.3)), and average (31.4). The other functions shown in this figure are needed for integrity verification and are described (at least briefly) in the following section.

31.3.3 Fault Monitoring

While the processing needed to generate pseudorange corrections from raw measurements is not demanding,

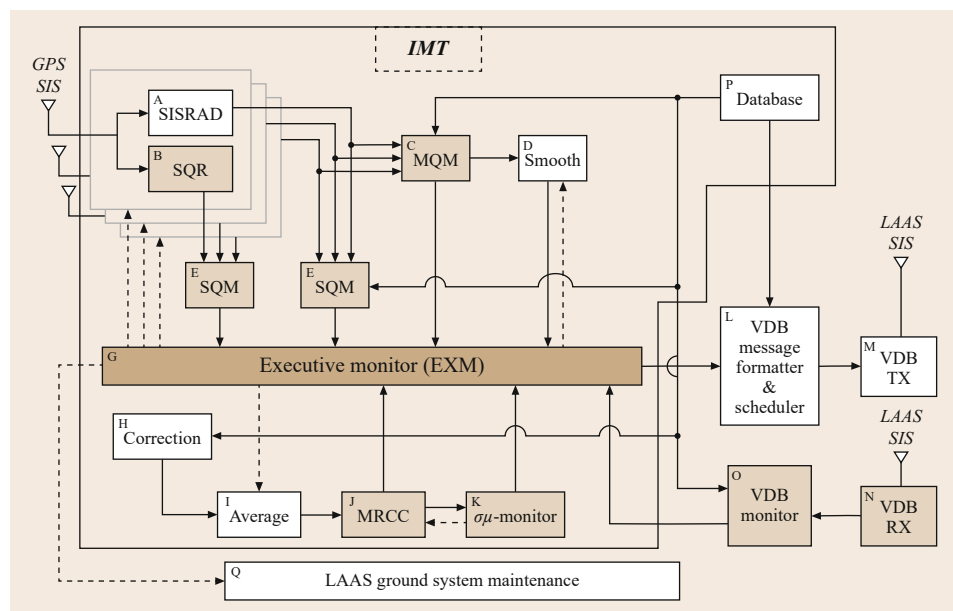


Fig. 31.3 Block diagram of GBAS ground station processing, integrity monitoring, and measurement exclusion logic

it is dwarfed by the many different algorithms used to verify the integrity of these measurements (and thus the broadcast corrections) in real time. Detailed descriptions of these algorithms are beyond the scope of this chapter. Summaries of the key algorithms and the anomalies that they address are provided along with references that contain further details.

The failures and/or anomaly conditions that could create hazardous errors for GBAS users if not detected or mitigated can be broken down into three categories:

- **Faults within GNSS:** This includes failures within GNSS satellites and failures or errors in the control segment that manages a given GNSS constellation, such as the GPS operational control segment (OCS).
- **Faults within augmentation system equipment:** This includes failures within individual GNSS reference receivers as well as anomalous multipath at reference receiver antennas.
- **Signal propagation anomalies:** This includes anomalous spatial gradients when GNSS signals pass through the ionosphere or the troposphere on their way to reference and user antennas. This can result in measured pseudoranges at the reference receiver that are significantly different than those measured by users. Other impacts, such as ionosphere *scintillation* (Chap. 39), are much more likely to cause loss of lock (or monitor detection) and thus lead to continuity breaches as opposed to loss of integrity.

The terminology of Fig. 31.3 is useful in further classifying the threats that the GBAS ground system must guard against. The following functions in that figure protect against the following subclasses of failures or anomalies.

Signal Deformation Monitoring (SDM)

This examines both the received power of the incoming satellite signals and the quality of the L1 C/A-code waveforms, meaning the degree to which the transmitted waveforms match the ideal ones. Significant waveform anomalies have occurred at least once, on GPS space vehicle number/pseudo-random noise (SVN/PRN) 19 as discovered in 1993, and nominal waveforms have small imperfections that must be accounted for in the design of GBAS [31.26, 27]. While most waveform abnormalities are highly correlated between reference and user receivers, these receivers are not required to be identical; thus large anomalies can cause significant user errors [31.28]. A threat model describing and bounding anomalous signal deformation for existing GPS satellites on L1 was developed based on a detailed analysis of the SVN/PRN 19 event [31.26, 28] and is used to quantify the integrity threat to GBAS from signal deformation after the implementation of the monitoring described below.

Detection of waveform abnormalities requires receivers that output GNSS measurements at multiple points along the code correlation peak so that asymmetries and unusual features can be observed and identified [31.28, 29]. This is performed by the *SQR* or *signal quality reception* function in Fig. 31.3. In prac-

tice, to avoid hardware duplication, this *SQR* capability is built into the reference receivers that also provide the measurements from which pseudorange corrections are generated [31.30].

Figure 31.4 shows an (exaggerated) example of a deformed code correlation peak and illustrates the monitor test statistics that are defined to detect significant deformations. The *ideal* code correlation peak differs slightly with pseudorange noise (PRN) code but is basically a perfectly symmetric triangle with *rounding* at the top due to the limited RF bandwidth in the GPS receiver. Multipath deforms the trailing (right-hand) side of the correlation peak, while satellite-generated signal deformation can alter both sides. The *delta* and *ratio* metrics shown in Fig. 31.4 quantify the difference between the observed correlation peak for each satellite (provided by the *SQR* function) and the ideal peak for the PRN broadcast by that satellite [31.28]. Thresholds on these metrics (to identify actual SDM faults within the continuity budget) must consider nominal deformations due to multipath and the nominal signal deformations mentioned above.

SDM is the major component of *signal quality monitoring*, or *SQM*, which is shown within Fig. 31.1. Additional monitoring that falls within this category includes testing the measured signal power from each satellite at each reference receiver to confirm that it falls within expected norms [31.31, 32]. This is not only a test of satellite performance but also a potential indicator of external RF interference (*RFI*), which will be discussed later.

Another satellite performance monitor with multiple purposes is that for code-carrier divergence, or *CCD*. CCD is caused by ionospheric delay that affects pseudorange and carrier measurements with equal magnitude but opposite sign. This degrades the accuracy of

carrier smoothing to an acceptable degree as long as the CCD rate (twice the rate of ionospheric delay change) is small, as it typically is. The CCD monitor in GBAS estimates the CCD rate in real time based on feeding the difference between raw L1 code and carrier-phase measurements into two successive smoothing filters, both with time constants of around 30 s. The resulting estimate of CCD rate is compared to a threshold, and a flag is issued if this rate exceeds an acceptable level [31.33]. In theory, it is possible for satellites to lose coherence between code and carrier elements of the broadcast signal and thus generate an unacceptable level of CCD. This is thought to be exceedingly unlikely, which means that, in practice, the CCD monitor serves as a means to detect unusual ionospheric behavior. The importance of this feature in mitigating threats due to large ionospheric spatial gradients will be explained later.

Data Quality Monitoring (DQM)

While this considers all of the navigation data broadcast by each satellite, its focus is the correctness of the ephemeris data that provides knowledge of the satellite position in space at any given time. Because GBAS reference receivers and aircraft are separated by no more than 10–20 km during the most critical flight phases, only very large ephemeris errors, such as a difference of over 1000 m between reported and actual satellite position, are potentially threatening. Errors so much larger than what is typical (differences of several meters) are considered to be possible from two causes [31.34].

Type A failure. OCS generation and upload of erroneous navigation data along with a maneuver of the affected satellite (i. e., the satellite changes orbits).

Type B failure. OCS generation and upload of erroneous navigation data without a maneuver of the affected satellite (i. e., the satellite remains in the same orbit).

Type B failures represent the simpler scenario where an observation, mathematical, or transcription error occurs in the process of generating updated ephemeris parameters for satellites that do not experience any external thrust that would cause an orbit change. GPS users would see a series of valid (correct) ephemeris messages (typically updated at 2 h intervals) followed by the sudden appearance of invalid (grossly erroneous) data after a particular message update. Therefore, consistency checking between old and new ephemerides is one means of detecting changes large enough to represent errors hazardous to GBAS. Because individual ephemeris messages are tightly fit to 2–4 h periods to maximize accuracy, and GPS satellite orbits change significantly (relative to the errors of con-

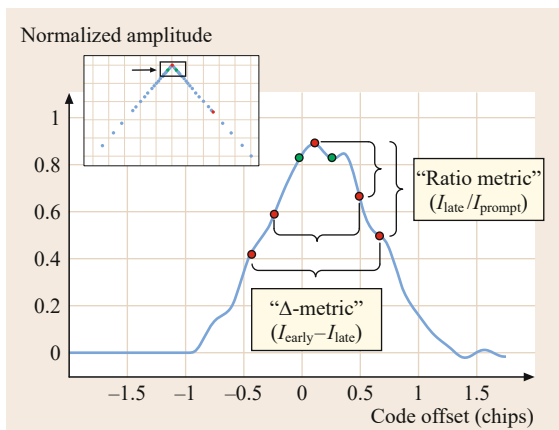


Fig. 31.4 Code correlation peak monitor test statistics for SDM

cern) over periods of 6–12 h, some degree of correcting for nominal satellite motion is desirable. The so-called *FOH-YETE* test described in [31.34] is a good compromise between simplicity and ability to detect significant type B failures. More involved orbit fitting of GPS satellites can greatly reduce the size of detectable errors, but the practical benefits of this additional complexity are small.

Type A events may also involve generation or upload mistakes but are complicated by the existence of a satellite maneuver. Deliberate satellite maneuvers are conducted occasionally to maintain the constellation and correct for normal orbit variation. Maneuvers change satellite orbits significantly; thus ephemeris parameters broadcast prior to a maneuver would be hazardous to use. To prevent this, satellites are flagged as *unhealthy* preceding maneuvers (meaning that their broadcast health bits indicate *do not use*) and are not flagged as *healthy* again until the maneuver is completed and new ephemeris messages reflecting the revised orbit are generated, uploaded, and double-checked for accuracy. Faults during this process that lead to hazardously incorrect ephemeris messages being broadcast by satellites flagged as healthy are denoted as type A failures.

Figure 31.5 shows an example of a relatively simple type A failure that occurred on GPS SVN 54 (PRN 18) on 10 April 2007 [31.35]. The plot shows the effects of this failure on uncorrected C/A-code range error of the affected satellite as measured at Honolulu, Hawaii along with the resulting 3-D position error at Honolulu, Los Angeles, California, and Billings, Montana (all in the western part of the US). The measured range error at Honolulu grew to about 350 m, and position error at Honolulu grew to over 500 m before the fault was noticed and the satellite uploaded to broadcast an *unhealthy* flag, which occurred about an hour after the fault began.

Subsequent investigation of this event showed that this was the result of a normal, planned orbit maneuver of SVN 54. The only problem was that the satellite was not updated to broadcast a status of *unhealthy* prior to beginning the maneuver. As a result, SVN 54 began moving away from its old broadcast ephemeris while still indicating its status as *healthy*. The fact that SVN 54 would be maneuvered was alerted ahead of time by a forecast Notice Advisory to Navstar Users (NANU) message, but NANU messages are not included in the broadcast navigation data and are not meant to substitute for the broadcast health status.

Type A faults of this sort are not directly observable to GBAS ground stations, but they can be readily detected by observing the computed pseudorange and carrier-phase corrections and changes over short inter-

vals (carrier-phase corrections are not normally broadcast but can be computed and used for monitoring – see [31.8]) [31.34]. In Fig. 31.5, the increasing range error experienced at Honolulu would have appeared as an increasing pseudorange correction to a GBAS ground station sited there and would have exceeded any reasonable value of this correction well before the satellite was flagged as *unhealthy* (and before any significant error was experienced by GBAS users). Demonstrating that all possible type A faults are protected by these checks requires extensive Monte Carlo simulations of possible orbit maneuvers in which conditions most difficult for these checks to detect are emphasized in the sampling of possible maneuvers [31.36].

Measurement Quality Monitoring (MQM)

This refers to monitoring of the time consistency of pseudorange and carrier measurements received at each ground reference receiver (or the average of all reference receivers). Sudden, sizeable inconsistencies in the received pseudorange, carrier-phase, or the difference between the raw pseudorange measured on a particular epoch and the pseudorange projected by the carrier-smoothing filter from the two most recent carrier-phase measurements (*carrier-smoothed-code innovations* – see (31.5) below) indicate the presence of a fault either in the received satellite signals or in one or more receivers tracking the affected satellites. To protect users, these must be detected quickly after they occur and either removed or corrected depending on their source and nature. Note that *cycle slips* in receiver carrier-phase tracking loops are an example of a receiver fault and are mitigated by MQM. Cycle slips must be sufficiently rare to meet the prior probability assumed by MQM, which places a lower bound on the signal power at which reliable carrier-phase measurements can be obtained in GBAS.

MQM algorithms for checking the time consistency of adjacent carrier-phase measurements and carrier-smoothed-code innovations are given in [31.25, 37]. The former is complicated by the need to duplicate the clock-adjustment steps shown (for computing pseudorange corrections) in (31.3) and (31.4). For each receiver at each epoch, clock-adjusted carrier-phase measurements for the last ten epochs (5 s) are used to fit a second-order polynomial that is then applied to predict the current (just acquired) phase measurement. The difference between actual and predicted phase measurement is compared to a threshold based both on expected noise of nominal carrier-phase measurements fit over 10 s and the minimum differences that could lead to threatening user errors. The *acceleration* and *velocity* values resulting from the second-order polynomial fit are also compared to their own thresholds.

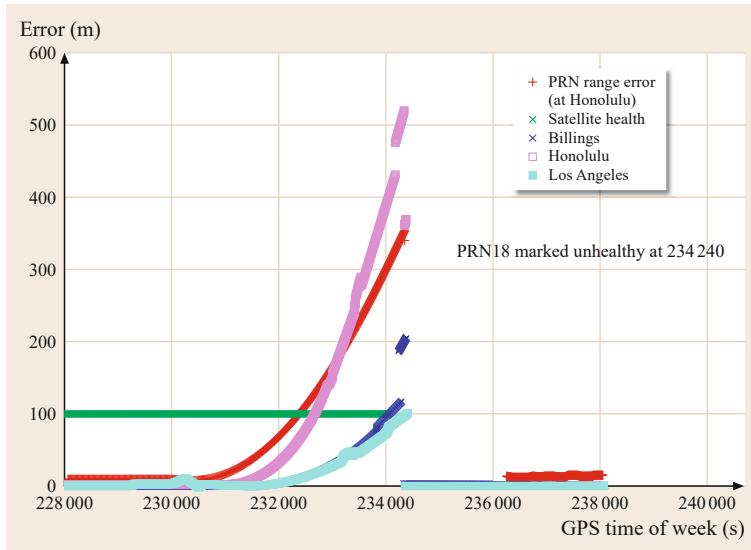


Fig. 31.5 Ephemeris type A failure example from April 2007. Standard positioning service (SPS) 3-D position error during PRN 18 anomaly

The monitor that detects anomalies in carrier-smoothed-code innovations is much simpler and can be expressed as follows using the notation of the smoothing filter definition in (31.1)

$$\text{Inno}(k) = \text{PR}_r(k) - [\text{PR}_s(k-1) + f(k) - f(k-1)] . \quad (31.5)$$

As with the carrier-phase consistency test, the innovation (Inno) derived from (31.5) is compared to a threshold that separates acceptable from unacceptable discrepancies.

Distinguishing between satellite and receiver faults detected by MQM is important but is typically left to executive monitoring (EXM), which will be described next. Unlike satellite faults, which normally lead to the exclusion of the affected satellite(s), faults that can be confidently limited to failures or cycle slips of a single reference receiver can often be *patched* by a variation of the algorithm that detects (and thus measures) the inconsistency between measurements. For the innovations test shown in (31.5), patching is performed by replacing the raw pseudorange measurement for this epoch with the one projected from the carrier-phase measurement update (presuming that this condition is limited to one reference receiver and did not also occur on the previous epoch). Patching of carrier-phase discrepancies is more complex but uses the same concept of replacing the (faulty) current measurement with the projection of what the measurement should have been from past measurements. If the code and/or carrier phase offsets resulting from a cycle slip can be patched with acceptable integrity, continuity is improved by preventing the loss of measurements from the affected reference receiver.

Detection and exclusion of satellite-driven faults detected by MQM, which are normally due to clock failures and are the most common GPS satellite anomalies, is important for two reasons. First, satellite faults that affect the time consistency of GPS measurements make it potentially unsafe to linearly extrapolate old pseudorange corrections forward several seconds in time using the range-rate correction (RRC), which is allowed when VDB messages are missed by the aircraft. For this reason, monitoring of unsafe levels of range acceleration (i. e., time-changing terms not included in the estimated range rate) is known as *excess acceleration monitoring*. Second, satellite behavior that falls well outside specified levels (e.g., those given in the *GPS SPS Performance Standard* [31.38]) indicates that the satellite is in an unhealthy mode, which means that the prior probability of failure assigned to healthy satellites (typically 10^{-5} per satellite per hour for GPS) no longer applies to this satellite [31.39]. This logic also applies to observed low satellite signal power under SQM, which may not be hazardous by itself but which invalidates the assumptions under which the safety of use of the affected satellite is guaranteed.

Executive Monitoring (EXM)

As mentioned above, EXM represents a series of logical functions and software execution paths that manage the flow of results from the monitors listed above (and those to follow). It determines which measurements are safe to use in calculating the broadcast pseudorange corrections and other integrity-sensitive values based on the combined monitor outputs, usually expressed as binary flags (pass-fail) for each measurement tested by each monitor. Because integrity must be protected to very

low probabilities, measurements made *questionable* by the combined monitor outputs may be discarded even if they are not individually *flagged* (result in failed tests) by any monitor.

Figure 31.3 helps to illustrate how EXM interacts with monitoring and measurement processing in the Stanford Integrity Monitor Testbed developed to show the feasibility of GBAS ground system monitoring [31.31, 37]. Here, while EXM activities occur throughout the processing of each epoch of measurements, the key logical elements of EXM are divided into two phases. In each epoch, the first phase occurs once the calculations and monitors shown above the EXM box in Fig. 31.3 are completed, which includes the monitors discussed to this point. Each of these monitors produces one *pass-fail* flag for each receiver channel, meaning each visible satellite tracked on each operating reference receiver. For example, if four reference receivers are tracking ten satellites, a total of 40 channels are tracked, and each channel has multiple *pass-fail* flags from each of the monitors executed to this point. These monitor flags are *OR*-combined such that a given channel receives a *fail* flag if any of its tests produce a *fail* result. From this point, a series of logical *measurement isolation cases* are applied as described in [31.31].

In most cases, measurements excluded by EXM enter *self-recovery* mode, in which the smoothing filters for the affected channels are restarted, and the measurements are retested with much tighter thresholds to see if any evidence of a fault remains. Tighter thresholds are used for recovery because excluded measurements are presumed to still be faulted until proven otherwise, whereas unexcluded measurements assume a low prior probability of failure before exclusion. If self recovery is unsuccessful after two or three attempted restarts, the measurements enter *external maintenance* mode, meaning that external maintenance is needed to restore them to service. Under certain threatening circumstances, including the case where all ground system measurements enter *external maintenance* mode, an *alarm* is issued that terminates service from that GBAS facility until external maintenance arrives to check and restore the system to full working order.

A *second phase* of EXM applies to the monitors derived from *B-value* statistics that compare measurements across reference receivers, as described under *multiple receiver consistency check* below. The same logical concepts apply, but the process of exclusion is iterative and includes multiple steps because the inclusion of receiver clock adjustments from this point forward affect all corrections and B-values.

Note that there are many ways to integrate EXM within the processing structure of a GBAS ground system. The method outlined above focuses on several

steps of logic resolution to remove all threatening measurements. Another approach is to specifically arrange the order of monitor execution so that each monitor acts as a *gate* on the measurements. In other words, all measurements that pass the previous monitors are fed into the next monitor, and only those that pass that monitor get propagated onward. The general order of Fig. 31.3 is followed, in that monitors that are aimed at detecting satellite failures precede those that are aimed at detecting receiver faults. However, this successive-gating procedure is insufficient by itself. Logical steps and cross-checks among multiple monitors need to be added to this *order-based* approach to assure that all potentially hazardous measurements are removed.

Multiple Receiver Consistency Check (MRCC)

Once the first phase of monitor tests and EXM resolutions described above has been completed, the surviving measurements are used in equations (31.2), (31.3), and (31.4) to calculate receiver clock adjustments and pseudorange corrections for each satellite. Reference receiver redundancy is now exploited to cross-check the smoothed and clock-adjusted candidate corrections from each receiver. For each satellite n with a computed correction $PRC(n)$ from (31.4), *B-values* are calculated as follows for all receivers who contributed to that correction [31.24]

$$B_{PR}(n, m) = PR_{corr}(n) - \frac{1}{M(n) - 1} \sum_{\substack{i \in S_n \\ i \neq m}} PR_{sca}(n, i). \quad (31.6)$$

Each $B_{PR}(n, m)$ (or $B_{n,m}$) value represents the error in the resulting pseudorange correction on satellite n that would occur if the measurement from receiver m were faulty. Here, faulty means that the measurement is not bounded by the zero-mean Gaussian distribution assumed for it under nominal conditions (more on this to follow). If receiver m were faulty, the *true* correction would be given by the average of the measurements from the other (nonfaulted) receivers tracking satellite n . Since the faulted measurement from receiver m is in fact included in the average used to generate the correction, the value $B_{i,j}$ indicates the error induced by this.

Given this logic, large values of $B_{n,m}$ suggest that the channel on receiver m tracking satellite n is faulted, and it is the role of MRCC and its associated EXM to exclude these channels before the final pseudorange correction is generated and broadcast. This starts by identifying (*flagging*) the B-values of any channels that exceed a preset threshold based on the bounding zero-mean distribution of B-values under nominal conditions. If any channels are flagged, at least one must

be removed, which means that the common satellite set and thus the clock-adjustment values for each reference receiver will change according to (31.3), leading to revised corrections and B-values.

Because it is possible (and indeed likely) for multiple B-value *failure* flags to be generated by a large-magnitude fault on a single measurement that propagates to other channels via the clock adjustment, EXM uses a special procedure of first excluding the *largest* fault (meaning the channel whose B-value exceeds its threshold by the highest percentage). This exclusion results in new corrections and B-values that are rechecked by MRCC. If flags remain on the (re-computed) B-values, logical isolation steps targeted at individual channel flags based on the EXM isolation cases shown above are applied in an iterative manner, meaning that the process of clock adjustment, correction, B-value recalculation, and threshold checking is repeated after each set of trial exclusions. This process should terminate in a viable, valid set of broadcast corrections and B-values within 1–3 iterations. If not, further exclusion iterations are abandoned, and the ground station is unable to broadcast any valid pseudorange corrections for this epoch (*self recovery* begins on all channels) [31.31].

Sigma-Mu Monitor

Receiver faults detected by MRCC as described above are rare for well-sited ground receivers. When they occur, the presence of large and unusual multipath from nearby reflecting sources is one likely cause (in addition to internal failures of the receiver, antenna, and connection hardware). MRCC and EXM remove faults that are potentially hazardous right away, but an additional concern is that unusual multipath is not immediately hazardous but is sufficient to violate the error bounds computed by GBAS users (to be discussed later). These more subtle receiver faults may be detected by MRCC given enough time, but additional monitors of the statistics of ground station receiver errors are added here to attempt to detect these conditions sooner: within hours for significant error increases or days for smaller ones.

Both the mean and the sigma of reference receiver errors are estimated using the just-computed (and passed by MRCC) B-values as inputs. Standard statistical estimation of the mean and sigma of a random process is used as one monitor, and this will normally detect significant error increases within hours to days. However, detection of larger errors (those just small enough not to be reliably detected by per-epoch MRCC) in minutes to several hours is desired, and this can be achieved by one or more additional approaches adapted from statistical quality control techniques. One is cumulative sum or *CUSUM* filtering, as described

in detail in [31.40]. Exponential (weighted toward the present) moving averages is another standard technique that achieves similar results. A simplification of the CUSUM technique is to accumulate the number of B-value threshold exceedances over time (using multiple thresholds below the MRCC threshold) and use these statistics to alert to significant error increases.

For well-sited ground systems with a significant performance margin built into them, flags from any of these monitors should be very rare. If they occur, the same iterative EXM procedure used to resolve MRCC flags can be used.

Message Field Range Check (MFRC)

The last step before the computed pseudorange corrections and correction rates can be approved for broadcast is to confirm that they fall within the message sizes that can be transmitted to users according to the ICD definition of GBAS message type one (this also applies to any broadcast parameters that change in real time based on reference receiver measurements). These limits are ± 327.67 m for the pseudorange correction (PRC) and ± 32.767 m for the correction rate (RRC) [31.8]. In practice, preliminary versions of these values are checked in advance (with tighter thresholds) as part of DQM monitoring to protect against type A ephemeris faults as described above. Therefore, this second check just before the values are sent to the VDB transmitter represents a *sanity check*.

31.3.4 User Processing and Integrity Verification

GBAS users who receive differential corrections and integrity information from it apply the broadcast information to improve their accuracy and guarantee their integrity in several steps [31.9]. First, the GPS satellite measurements of multiple redundant airborne receivers are individually checked for received signal power, properly decoded navigation data, CCD (for Category II/III users), and other basic quality tests. (Note that, unlike the ground system, the outputs of multiple airborne receivers are treated independently with regard to GBAS and are not compared together until later in the chain of airborne navigation processing, after GBAS position outputs are provided from each receiver). Measurements deemed acceptable for use then have the broadcast pseudorange corrections applied to them. The basic equation for the application of corrections is [31.9]

$$\begin{aligned} \text{PR}_{\text{corr_air}}(n) = & \text{PR}_{\text{air}}(n) + \text{PR}_{\text{corr}}(n) \\ & + \text{RR}_{\text{corr}}(n)(t - t_{z_count}) \\ & - \text{TC} + c(\Delta t(n))_{L1} \end{aligned} \quad (31.7)$$

with:

- $PR_{\text{corr_air}}(n)$ is the resulting corrected smoothed pseudorange airborne measurement for satellite n and the current time t
- $PR_{\text{air}}(n)$ is the smoothed pseudorange airborne measurement for satellite n and time t before applying the correction
- $PR_{\text{corr}}(n)$ is the pseudorange correction for satellite n broadcast by the ground system (31.4) that is valid at time t_{z_count} , which is also included in the broadcast message
- $RR_{\text{corr}}(n)$ is the range rate correction for satellite n broadcast by the ground system that is valid at time t_{z_count}
- TC is the tropospheric delay correction computed according to [31.9]
- c is the speed of light in vacuum
- $(\Delta t(n))_{L1}$ is the L1 clock correction from the navigation message for satellite n applied at time t .

Several protocols apply to the use of GBAS corrections [31.9]. The most important is that only satellites for which corrections are broadcast can be applied to compute GBAS user locations. It is not uncommon for an airborne receiver to track more satellites and obtain more valid measurements than the ground system provides corrections for. The primary reason for this is that the ground system reference receiver antennas cannot reliably track down to elevation angles as low as those that are visible from aircraft. Note that this protocol allows the ground system to protect users from threats detected on the ground – excluding the affected measurements means that no corrections will be generated for them, and users will also have to discard them. Users also confirm that, for each satellite corrected with ground measurements, the received clock and ephemeris navigation data at the user matches what was used by the ground system to compute the corrections. The ground supports this by broadcasting the 16 bit encoded cyclic redundancy check (CRC) of its received data for each approved satellite, allowing the airborne system to execute the same CRC algorithm on its received data and confirm that there is a match [31.8].

User position (in three dimensions: along-ground track, cross-track, and up) and time is computed first by defining a measurement weighting matrix \mathbf{W} , which is the inverse of an $N \times N$ diagonal matrix (N is the number of usable satellites) with entries given by the total (fault-free) error variances (σ_i^2) of each satellite i . This total variance is the root-sum-square (RSS) of ground and airborne noise components as follows [31.9]

$$\sigma_i^2 = \sigma_{\text{pr_gnd}}^2[i] + \sigma_{\text{tropo}}^2[i] + \sigma_{\text{pr_air}}^2[i] + \sigma_{\text{iono}}^2[i]. \quad (31.8)$$

Here $\sigma_{\text{pr_gnd}}$ bounds the ground pseudorange measurement error after carrier smoothing. It is a key parameter broadcast for each satellite in message type 1 and is based on the known azimuth and elevation angles of that satellite from the ground station, data collected at the ground station, and multipath error models [31.8]. Demonstrating that the broadcast values of $\sigma_{\text{pr_gnd}}$ are sufficient bounds on actual rare-event errors is a significant challenge for ground system design – see [31.41–43].

$\sigma_{\text{pr_air}}$ bounds the airborne pseudorange measurement error after carrier smoothing. It is determined by each user and is bounded by the airborne accuracy designator A or B models given in [31.9].

σ_{tropo} bounds the impact of tropospheric decorrelation due to altitude offset between user and ground station. It is determined by the user-estimated height offset and a sigma parameter broadcast by the ground in message type 2 [31.8, 9].

σ_{iono} , finally, bounds the impact of ionospheric decorrelation due to horizontal separation between user and ground station. It is determined by user-estimated horizontal separation and velocity estimates and a $\sigma_{\text{vert_iono_gradient}}$ (or σ_{vig}) parameter broadcast by the ground in message type 2 [31.8, 9]. The broadcast value of σ_{vig} may be inflated to help mitigate rare, very large spatial gradient events that cannot be bounded by the broadcast σ_{vig} . This is described further in Sect. 31.3.5 to follow.

The standard equation to solve for the four position states (vector \mathbf{x}) from the N smoothed and corrected measurements (vector \mathbf{y}) then applies (this ignores the successive steps of linearization needed to make this model valid [31.9])

$$\mathbf{x} = (\mathbf{G}^T \mathbf{W} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{W} \mathbf{y}, \quad (31.9)$$

where \mathbf{G} is the standard $N \times 4$ geometry matrix in which each row i ($i = 1, \dots, N$) is defined by the unit vectors from the user to satellite i

$$\mathbf{G}_i = \begin{bmatrix} -\cos \text{El}_i \cos \text{Az}_i \\ -\cos \text{El}_i \sin \text{Az}_i \\ -\sin \text{El}_i \\ 1 \end{bmatrix}^T. \quad (31.10)$$

User integrity is provided first by requiring users to exclude satellites for which corrections are not broadcast and is verified (or quantified) in real time by the user's calculation of *protection levels* in the position domain. These represent bounds on user position error at specific (very low) probabilities suballocated from the required loss-of-integrity probabilities for GBAS flight applications, including Category I/II/III precision approach. Protection levels are unique to each GBAS user

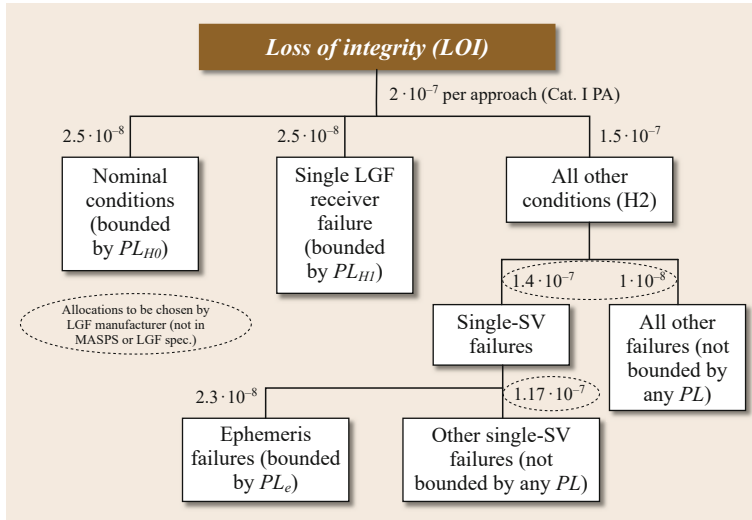


Fig. 31.6 GBAS top-level integrity risk allocation for Category I precision approach

because they include user-specific error terms, for example $\sigma_{pr_air}^2$ in (31.8), and include only the intersection of the set of satellites tracked by ground and airborne receivers. By comparing protection levels computed in real time to the safe error limits (*alert limits*) for a given type of operation, they allow each user to determine if continuing the operation is safe or if the operation should be aborted.

In simple terms, protection levels are constructed by extrapolating bounding probability distributions of expected user errors to the suballocated integrity probabilities mentioned above. Multiple protection levels are defined as needed to account for specific failure conditions, and the final protection level applied by users is the maximum of the protection levels defined to cover specific conditions. Figure 31.6 illustrates how the overall integrity requirement for Category I (GBAS Approach Service Type GAST-C) GBAS is suballocated to a set of scenarios, some of which are assigned protection levels. The top-level allowed integrity risk from Table 31.1 is $2 \cdot 10^{-7}$ per approach. This is first subdivided into three categories: *H0*, *H1*, and *H2*. The first two of these are defined below and are covered by protection levels. The third, *H2*, represents all other conditions (not *H0* or *H1*) and includes satellite faults and atmospheric anomalies, which is why it receives the bulk (75%) of the integrity allocation. Within the *H2* category, faults of single satellites are separated out, and among these, ephemeris faults are distinguished and given a unique suballocation because they are covered by a specific protection level. The other single-satellite faults, along with multiple-satellite faults and all other faults not yet described, do not have protection levels assigned to them. Each ground-system manufacturer can allocate probabilities among the faults not spec-

ified to have protection levels, which provides some flexibility.

The simplest protection level is that which applies to nominal conditions, called the *H0 case* to represent the *default hypothesis* in GBAS. Nominal conditions mean that a zero-mean Gaussian distribution with the total (ground plus airborne) variance computed in equation (31.1) bounds the actual distribution of range errors to the required integrity probability. If this is the case and range-domain error bounding persists in the position domain (e.g., errors among different satellites are uncorrelated), the *H0* protection level for the vertical direction is [31.9]

$$VPL_{Apr_H0} = K_{ffmd} \sqrt{\sum_{i=1}^N s_{Apr_vert,i}^2 \sigma_i^2 + D_V}, \quad (31.11)$$

where $s_{Apr_vert,i}$ is the vertical component for satellite *i* from the matrix **S**, which is the pseudo-inverse of the geometry matrix **G** and is computed from (31.9)

$$\mathbf{S} = (\mathbf{G}^T \mathbf{W} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{W}.$$

K_{ffmd} is a scalar multiplier that represents the required integrity risk under the fault-free scenario (it is about 5.8 for Category I GBAS), and D_V is the difference in vertical position between 30 s-smoothed and 100 s-smoothed pseudorange solutions. Only 100 s smoothing is used in GAST-C for Category I, so $D_V = 0$. In GAST-D, 30 s smoothing is used in both ground and airborne positioning, but much of ground monitoring is still based on 100 s-smoothed pseudoranges. D_V represents the position-domain difference of the two different smoothing times, which is sensitive to anomalous ionospheric divergence. Thus, (31.11) propagates

the bounding range domain error distribution into the position domain, extrapolates it to the appropriate sub-allocated integrity probability (via K_{fmd}), and adds an estimate of ionospheric divergence in the position domain (D_V) for Category II/III approaches.

Two additional protection levels are defined that cover two specific failure conditions. The first is a failure of a single reference receiver that was not detected and excluded by MRCC within the ground system. This is known the *H1* case and has the following protection level [31.9]

$$\text{VPL}_{\text{Apr_H1}} = \max(\text{VPL}_{\text{Apr_H1}}[j]) + D_V, \quad (31.12)$$

where

$$\text{VPL}_{\text{Apr_H1}}[j] = |B_{j,\text{Apr_vert}}| + K_{\text{md}}\sigma_{\text{Apr_vert_H1}} \quad (31.13)$$

$$B_{j,\text{Apr_vert}} = \sum_{i=1}^N s_{\text{Apr_vert},i} B_{ij}. \quad (31.14)$$

Here, the B-values come from the broadcast products of MRCC as described before for satellite i and reference receiver j , while K_{md} is the scalar multiplier that represents the missed-detection probability needed to cover the *H1* fault scenario. As in MRCC, the B-values express the actual correction errors that would occur under each *H1* fault subhypothesis, and *H1* protection levels include this as a bias (converted to the position domain) that is added to nominal errors extrapolated to a suballocated probability for *H1* events after correction for the relative rarity of these events. To be used in this way, B-values must be for each ground receiver (up to $M = 4$ receivers) and satellite must be included in the broadcast message type 2 [31.8]. The K_{md} values given in [31.9] assume a per-receiver failure value (for these purposes) of approximately 10^{-5} per approach; thus this value becomes a requirement on ground system design (in practice, continuity requirements on reference receiver failures are more constraining).

Note that $\sigma_{\text{Apr_vert_H1}}$ and $B_{j,\text{Apr_vert}}$ in (31.13) have already been converted into the position domain using the same approach as applied to (range-domain) σ_i in (31.11). The range domain version of σ_{H1} is slightly higher than σ_i because the *H1* scenario has one fewer nominal reference receiver that contributes to the nominal vertical error distribution. In fact, (31.13) can be interpreted as the sum of the nominal error distribution from the nonfaulted measurements plus the bias due to the faulted measurement. In the *H1* scenario, this bias results from a (hypothetical) single reference receiver fault as measured by the B-values computed on the ground, but the same approach applies to all protection levels computed for fault hypotheses.

Note that one value of VPL_{H1} is computed for each reference receiver j (from 1 to M) in (31.13), and the maximum combined B-value over all M reference receivers is applied in (31.13). This follows from the general approach of computing individual values for each scenario and taking the maximum (rather than a weighted average) to cover all of them.

The second failure condition with its own protection level models undetected satellite ephemeris faults [31.9]

$$\text{VBP}_{\text{Apr_e}} = \max(\text{VBP}_{\text{Apr_e}}[k]) + D_V, \quad (31.15)$$

where

$$\begin{aligned} \text{VBP}_{\text{Apr_e}}[k] = & |s_{\text{Apr_vert},k}| x_{\text{air}} P_{k_x} \\ & + K_{\text{md_e_x}} \sqrt{\sum_{i=1}^N s_{\text{Apr_vert},i}^2 \sigma_i^2}. \end{aligned} \quad (31.16)$$

Here, $K_{\text{md_e_x}}$ is the scalar multiplier that represents the missed-detection probability needed to cover the ephemeris fault scenario, and x_{air} is the horizontal distance between user and reference station (the geographic centroid of the reference receiver antennas is broadcast to users for this and other purposes). P_{k_x} is a broadcast parameter that expresses the ground system's ephemeris detection capability in terms of the smallest spatial error gradient (for satellite k) that is assured to be detected with the required integrity probability. Therefore, the user remains potentially vulnerable to ephemeris faults that cause spatial error gradients at or below this level, and these must be accounted for by the ephemeris protection level. The inclusion of x_{air} allows each user to apply its own distance from the ground station, which is important for a fault whose user impact increases linearly with separation.

Analogous equations are used in real time to generate lateral protection levels (LPLs), meaning protection levels in the lateral (cross-track) direction. During precision approach, the requirements on vertical protection level (VPL) are almost always more constraining and are thus the focus of offline requirements analysis.

Protection levels resulting from failure modes other than the two for which specific protection levels exist (e.g., excess acceleration from satellite clock, signal deformation, etc.) must be covered by the three protection levels that are defined. In practice, since the equations for *H1* and ephemeris faults are specific to those cases, the *H0* protection level must bound all of these events. Demonstrating that this is the case requires analysis and (in most cases) simulation of each fault type to generate the worst-case range-domain user errors resulting

from them and comparing these to the range-domain protection level implied by the *H0* scenario. *Worst-case* in this context means the failure subscenario that potentially causes the largest differential range error either without detection or with monitor detection within the ground system time-to-alert. A detailed methodology for addressing the results of failure simulations is given in [31.44].

31.3.5 Additional Threats: RF Interference and Ionosphere

The consequences of two additional threats have become apparent over time as GBAS prototype systems have been fielded in various places. Both of these threats were recognized and treated by the original design of GBAS ground and airborne equipment but have proven to be more severe than first thought.

RF Interference (RFI)

RFI, as described in Chap. 16 refers to the imposition of external RF signals onto received GPS signals at the ground station or users in a manner that either makes the GPS signals unusable or subtly makes them less accurate. Strong interference to GPS from individual transmitters has been discovered several times in the past and can prevent the use of GPS signals over wide areas [31.45]. These events are rare, but more common interference results from what are known as *personal privacy devices* or *PPDs*, which are low-power jammers designed to *protect the privacy* of their users by preventing the use of GPS for tracking and surveillance. They are most commonly fielded in vehicles and typically achieve an effective jamming range of meters to tens of meters. *personal privacy devices (PPDs)* are illegal to operate in most places but are relatively easy and inexpensive to obtain over the Internet. The variety of devices available for purchase and their illicit nature means that their actual performance is variable; thus some devices may affect GPS over several hundred meters rather than the much shorter distances needed to *protect* a single vehicle [31.46]. For this reason, they can be troublesome to GBAS ground stations sited near major roadways [31.47].

The potential impact of PPDs on GBAS ground stations became evident in 2009, when the newly-fielded Honeywell SLS-4000 LAAS ground facility (LGF) began initial operation at Newark International Airport (EWR) in New Jersey, USA. The layout of this ground system is shown in Fig. 31.7. The site chosen within the property assigned to Newark Airport was not ideal and was selected due to the lack of suitability of several other locations considered. As a result, the four

reference receivers fielded at Newark are in almost a straight line with separations of about 100 m to make multipath errors among the antennas as statistically independent as possible. All four antennas are within 200 m of both lanes of the (very busy) New Jersey Turnpike (I-95). Several times a day, vehicles with strong PPDs passing by the airport would first make GPS unusable at one receiver, then the next, until service was lost (no valid corrections), and recovery from this condition took a significant period. The need to protect the integrity of the ground station measurements led to measurements being affected by PPDs to be excluded due to violations of one or more monitors, including the low-signal-power monitor, before the jamming became strong enough to prevent the receivers from making measurements [31.47].

Once this threat was evaluated and better understood, both physical and software changes were made to the Newark LGF (and other GBAS sites) to minimize the impact of PPD jamming. Software refinements to the monitors affected by PPDs and the EXM that handled monitor flags were key in preventing shutdown from PPD events while still protecting integrity. This was done by conservatively modeling the threat (from offline test data and RFI observations at Newark [31.46]) to determine the worst-case impact of several variations of refined monitor logic. The result was a validated means of tolerating the exclusion of two *jammed* reference receivers (out of four) for a brief period while still providing safe (albeit somewhat degraded) pseudorange corrections from the two *unjammed* ones. Since vehicles on the nearby roadway are almost always moving at speed, those with strong PPDs affect the revised LGF software but without bringing it to the *shutdown* point.

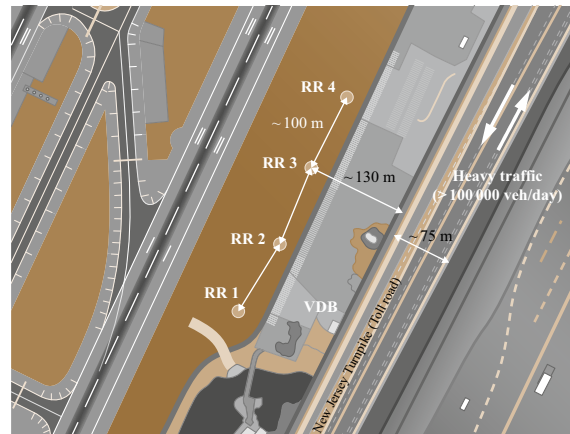


Fig. 31.7 GBAS ground system siting at Newark Airport, NJ, USA. Note four reference receiver antennas in a line close to a busy freeway

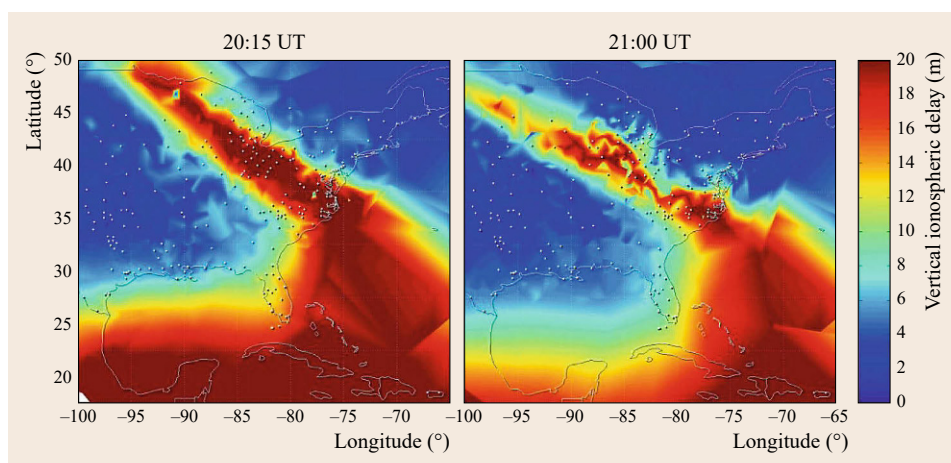


Fig. 31.8 Large-scale ionospheric disturbance over CONUS on 20 November 2003 (after [31.51])

This experience with PPDs at Newark, along with other experiences of unintentional jamming affecting GBAS ground stations and aircraft (e.g., see [31.48] for the problems caused by an indoor GPS signal repeater broadcasting strong signals visible to landing aircraft), suggests changes in the way that future GBAS ground systems are fielded. First, if at all possible (though not possible at Newark), reference receiver antennas should be placed further away (≈ 500 m or more) from roadways with significant traffic, and they should be spread out further to prevent a single relatively weak jammer from affecting more than one reference receiver at a time. Future ground system designs will support larger antenna separations, but the limited real estate available for siting at many airports will remain a significant constraint.

Ionospheric Spatial Decorrelation

Under normal conditions, the difference in ionospheric delay on smoothed pseudorange measurements made within 5–100 km of each other is a minor component of the total error budget for GBAS users and is conservatively bounded by the broadcast σ_{vig} term used to compute σ_{iono} in (31.8). In CONUS, the typical (1 s) spatial variation in zenith (i.e., vertical, at 90° elevation) ionospheric delay at the GPS L1 frequency is about 1 mm/km, and the value of σ_{vig} used to bound ionosphere is 4 mm/km to add margin to cover days with high levels of ionospheric activity [31.49]. However, during strong ionosphere storms or other severe anomalies within the ionosphere, this value can grow much larger and can threaten GBAS user safety. Because each GBAS ground station views the ionosphere from one location on Earth and does not normally share information with other locations, an ionospheric anomaly can arrive at the vicinity of a ground station and potentially cause hazardous errors before it would

be detected by GBAS ground and airborne monitoring.

The possibility of ionospheric spatial gradients large enough to be threatening was first discovered by analyses of Wide Area Augmentation System (WAAS) reference station data in 2002. In October and November of 2003, a very strong coronal mass ejection (CME) from the Sun created severe ionospheric anomalies that were noticed by WAAS and were analyzed after the fact by the continuously operating reference station (CORS) network of about 400 stations in CONUS at that time (now there are more than 1000 stations). Several years of WAAS and CORS observations in CONUS were collected and analyzed, with the result being a *threat model* that both describes the dynamics of these events from the GBAS viewpoint in simplified terms and provides bounding parameters on the variables in the model such that behavior outside the model is deemed to have negligible probability (see [31.50] for a general description of threat models and how they are used).

Figure 31.8 shows the ionospheric anomaly that generated the largest spatial gradients that have ever been discovered in CONUS or elsewhere in the middle latitudes (similar analyses have been done for Germany, Australia, and other locations). This occurred in the afternoon hours (local time) on 20 November 2003, during the period of coronal mass ejections mentioned above. In this figure, a zone of very high vertical ionosphere delay (in red) exists in between two zones of much lower delay (in blue). Thus, at the boundaries between the high and low-delay zones, very large spatial gradients occurred. The overall pattern of high delay moved roughly westward over time and eventually broke up into more irregular features, as shown in the plot for 21:00 Universal Time (UT) compared to 20:15 UT.

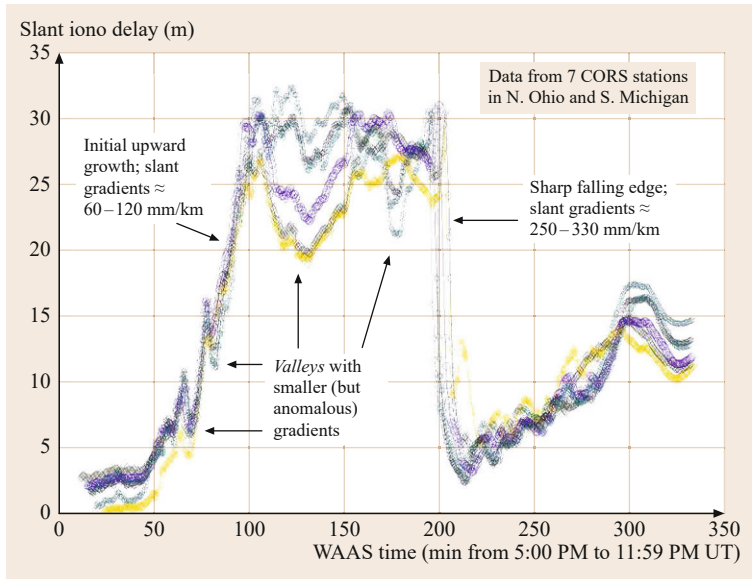


Fig. 31.9 Change in ionospheric delay observed on SVN 38 at seven adjacent stations in CONUS on 20 November 2003 (after [31.51])

Figure 31.9 examines the event shown in Fig. 31.8 by plotting the slant ionospheric delay over time as observed by seven CORS reference stations located close to each other in Northern Ohio and Southern Michigan that are tracking the same GPS satellite (space vehicle number SVN 38). This shows the rapid rise in delay caused by the leading edge of the high-delay zone as it passed over. The resulting spatial gradients between these stations were very high (60–120 mm/km) compared to typical levels of 1–5 mm/km [31.49]. The observed ionospheric delays remained anomalous and unsteady during the period when the high-delay zone affected the measurements. When the trailing edge of this zone reached the measurements a little before 21:00 UT, a sudden, very large drop in delay appears at all seven stations, creating spatial gradients between the stations of as large as 330 mm/km. This same event caused gradients as large as 412 mm/km between other pairs of stations in Northern Ohio.

Based on observations of this event and several others that also created anomalous gradients (but not this severe), a threat model covering what was observed (and validated to be due to ionospheric events) was developed and is shown in Fig. 31.10 [31.51, 52]. This threat model provides bounds on the parameters of a simplified physical model of an ionospheric spatial gradient affecting a GBAS station and nearby users. This model, shown in Fig. 31.11, assumes that the gradient is generated by a constant, linear change in delay between high and low (or low and high) levels, and this gradient wedge is moving with constant speed relative to the ground. Example numbers that fall within the threat model bounds are shown in Fig. 31.11, but any

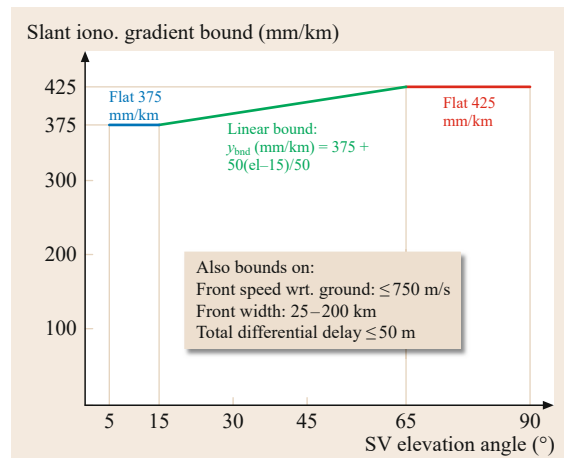


Fig. 31.10 Ionospheric gradient threat model established for CONUS (after [31.51])

numbers falling within these bounds are deemed to be possible (credible); thus user integrity must be demonstrated against all permutations of parameters within the threat space.

Figure 31.10 shows the most critical threat model parameter, which is the maximum slant ionospheric spatial gradient as a function of satellite elevation angle. Based on observed data, it varies from 375 mm/km at low elevation angles to 425 mm/km at high elevation angles. Note that the latter number bounds the observed largest event of 412 mm/km with a small amount of margin for measurement error [31.52]. Other bounds are also listed, including the gradient propagation speed relative to the ground (no greater than 750 m/s, but smaller

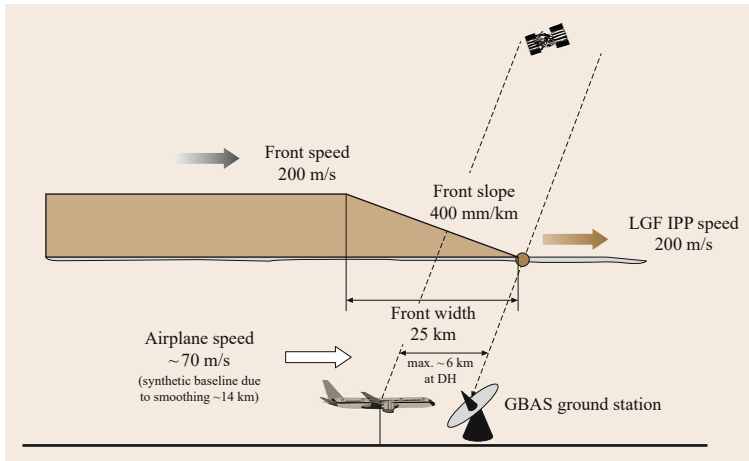


Fig. 31.11 Simplified linear wedge model of ionospheric spatial gradient (after [31.51])

speeds are harder to detect and thus more threatening), the width of the zone in which the delay is changing (no less than 25 km), and a limit on the total differential delay (width times gradient) of 50 m. The limit on total differential delay means that some permutations of gradient and width that would otherwise fall within the threat model are eliminated because their product exceeds 50 m and thus are considered noncredible.

While the CONUS threat model is thought to cover all mid-latitude GBAS sites, regions outside CONUS should conduct their own data analyses to confirm that no larger gradients exist. If only smaller gradients are found, a less-conservative threat model bounded by these lower gradients could be used. However, because of the rarity of the events that cause extreme gradients and the low likelihood of observing very large gradients in a small geographic area, the use of the CONUS threat model of Fig. 31.10 is recommended for all mid-latitude regions that do not discover gradients exceeding it [31.53].

When the CONUS threat model is implemented as shown in Fig. 31.11 and then applied to GBAS, it is quickly evident that the worst-case user differential range error is very large. The (simplified) equation that relates ionospheric spatial gradients to user differential range error is given by [31.9, 49]

$$E_{\text{iono}} = F_{\text{PP}} \times g_v \times (x_{\text{air}} + 2tv_{\text{air}}), \quad (31.17)$$

where E_{iono} is the resulting differential pseudorange error, g_v is the ionospheric spatial gradient in vertical (zenith) terms, t is the smoothing filter time constant (100 s for GAST-C GBAS), x_{air} is the ground-to-user separation, and v_{air} is the horizontal aircraft approach velocity (i.e., velocity of approaching the airport and the GBAS ground system). F_{PP} is the standard ionospheric obliquity factor that converts between zenith

and slant delay, but it is not needed here because gradients in the threat model are already expressed in *slant* (g_s) terms, where $g_s = F_{\text{PP}} \times g_v$.

Applying the maximum gradient ($g_s = 425 \text{ mm/km}$) with possible values of x_{air} (6 km for very large airports) and v_{air} (70 m/s is typical of jet aircraft) results in a (worst-case) differential error of 8.5 m on a single satellite, which is very large and would be difficult to handle with the existing VPL equations because these are compared to an alert limit of 10 m in the position domain (not the range domain). The vast majority of gradients approaching this size would be detected by one or more of the ground station monitors described in Sect. 31.3.3, such as the code-carrier divergence (CCD) monitor, that are sensitive to the time rate of change of ionospheric delay. However, it is possible for a gradient moving with respect to the ground to be counteracted by the motion of the satellite with respect to the ground, resulting in very little change of delay with time and thus very little to trigger the existing ground monitors.

Since very large differential range errors are possible and are not guaranteed to be detected by the ground station, a different technique for demonstrating the safety of Category I precision approaches under worst-case ionospheric conditions was needed. Two things resulted:

1. A relaxed requirement on the tolerable error limit (called TEL) based on the obstacle clearance surface (OCS) definitions for Category I precision approaches [31.54]
2. A new ground system integrity algorithm known as *geometry screening* that strategically inflates the broadcast VPL parameters (including the ephemeris decorrelation parameter P , σ_{vig} , and/or $\sigma_{\text{pr_gnd}}$) to ensure that any satellite geometries that could violate TEL are unavailable and thus cannot be used.

The procedures used to implement geometry screening in real time within GBAS ground systems are complicated and difficult to describe in simple terms. Detailed descriptions of two different approaches to the task are included in [31.55, 56]. The following summarizes the key steps in geometry screening:

1. Using the known set of approved satellites now and in the near future (e.g., the next 10 min), derive the set of possible user satellite geometries. This requires some assumptions on which subsets of satellites have been approved by the ground but may have been discarded (or unusable) by users (presumed to be approaching aircraft).
2. Using a simulation of the model in Fig. 31.11 with all possible satellite geometries, user approach geometries, and threat model parameters, determine the largest possible ionospheric-induced error in vertical user position (denoted as *MIEV*) for each possible user satellite geometry that would be deemed available by users (because the user VPL is below the normal Category I alert limit of 10 m). Note that *MIEV* also depends on the assumption of how many satellites can simultaneously be affected by the worst-case gradient (all combinations of two satellites are normally tested).
3. If *MIEV* exceeds TEL (about 28 m at the Category I approach threshold) for any possible user geometries, implement broadcast parameter inflation using trial increases of P , σ_{vig} , and/or $\sigma_{\text{pr_gnd}}$. Continue increasing a subset of these parameters until all possible user geometries become unavailable (VPL now exceeds the alert limit given the inflated broadcast parameters).

Implemented in this manner, geometry screening succeeds in protecting Category I GBAS users against worst-case ionospheric spatial gradients. Because of the conservatism involved in protecting each and every permutation of the threat combined with every possible user geometry, the result is significantly lower user availability. If geometry screening were not needed, GBAS Category I user availability would easily exceed 99% and might approach 99.9% at some airports. With geometry screening implemented based on the CONUS threat model and testing of all combinations of two satellites impacted at the same time, availability can fall below 99% for ground-to-approach-threshold separations of 1–3 km and below 98% for larger separations.

Given the lessons learned from GAST-C, the design of GAST-D GBAS to support Category II/III approaches implements several improvements to reduce the impact of anomalous ionospheric spatial gradients and improve the effectiveness of ground and airborne

monitoring [31.57]. First, as mentioned above, GAST-D positioning is based on a 30 s smoothing time constant, which significantly reduces the possible differential range error per (31.17). Second, CCD monitoring is required in user equipment as well as within the ground system. This is advantageous because airborne users are moving relative to the ground and thus can detect gradients that the ground might be blinded to (and vice versa). Combined with the use of D_V in the VPL calculations, this substantially increases the likelihood of detecting ionospheric divergence caused by anomalous spatial gradients.

Finally, GAST-D ground systems are required to implement new spatial gradient monitoring so as to guarantee that very large gradients that affect ground receiver will be detected regardless of any temporal *blinding* effects. To achieve this, differential carrier-phase test statistics between reference receiver antennas carefully sited to provide specific separations and directions of observability are needed. Several related algorithms have been proposed and are collectively called ionospheric gradient monitoring, or IGM [31.58, 59]. IGM is difficult to implement due to siting restrictions (see Sect. 31.3.6), small variations in reference receiver antenna errors that must be calibrated out to the extent possible, and the fact that even very large gradients only create small signatures over the short baselines used by IGM.

However, once IGM is implemented, it allows a large relaxation of the geometry screening required, and the remaining screening is done by each aircraft instead of within the ground system. Since each aircraft knows its own satellite geometry, the ground system need not cover all possible airborne geometries. That, combined with the reduced level of gradient that can escape undetected by IGM, leads to a significantly lower impact of geometry screening on precision-approach availability.

31.3.6 Equipment and Siting Considerations

Several aspects of ground system siting have already been discussed in previous sections. Siting focuses on choosing the best locations for the reference receiver antennas (to support monitoring while minimizing errors) and the VDB transmitter antenna (to cover the airborne user region with adequate, but not too strong, signal levels [31.8]). As noted above, reference receiver antenna siting follows these principles:

1. Where possible, separate the antennas by at least 100 m to minimize multipath error correlation among reference receivers, which weakens the ability of MRCC to mitigate individual receiver errors

- (the residual correlation at each site should be estimated and accounted for).
- Where short-baseline ionospheric gradient monitoring (IGM) is implemented to support GAST-D, antenna separations on the order of 150–300 m are needed [31.58, 59].
 - Where RF interference from nearby roadways or other sources is a threat, separate the antennas further to the degree possible to minimize the possibility of multiple receivers being affected simultaneously by the same interferer.

In addition, reference receiver antennas need to be located a sufficient distance away from nearby obstructions as to limit multipath errors to the bounding error models deemed acceptable from offline analysis. From years of trial-and-error and multipath error observations at several airports, siting guidelines to achieve this have been developed [31.60]. The concept of one or more *clear zones* that are free of significant multipath reflectors comes from the siting of ILS transmitters. An inner *clear zone* of perhaps 100–200 m in radius must be clear of all obstructions, including aircraft or other vehicles, while relatively small reflectors may be allowed in larger outer zones.

These guidelines take advantage of the use of multipath-limiting antennas, or MLAs, in GBAS ground systems. These devices go beyond normal GNSS antennas to establish a very strong roll-off between the gain pattern at low but positive elevation (where low-elevation satellites should be tracked) and low but negative elevation (which is the source of ground multipath). The first generation of MLA used in GBAS was a dual-element antenna with a helibowl facing toward zenith and a multi-element dipole for lower elevations that established the required gain roll-off. These two elements required two different receivers and *cross-over* software logic to combine their outputs into one *virtual* receiver for downstream GBAS processing [31.61]. The more-recent generation of MLAs uses a single multi-element dipole to track both high-elevation and low-elevation satellites [31.62], avoiding the need for two receivers per antenna and the resulting cross-over calibration. While not perfect, these antennas work very well in reducing both specular (single-reflection) and diffuse (multireflection) multipath to small levels that can be empirically measured and bounded by the broadcast $\sigma_{\text{pr_gnd}}$ as long as the above siting guidelines are followed.

One major advantage of GBAS siting with respect to ILS is that GBAS sites do not have to occupy specific locations near runway ends and that one GBAS ground station can support all runway ends. This benefit is limited by the need to find multiple *clear zones*

and by the need to limit the separation between the geographic centroid of the reference receiver antennas and the landing thresholds of the multiple runways being covered. The effects of anomalous ionospheric spatial decorrelation, in particular, get worse with increased ground-to-user separation. While the *geometry screening* used to mitigate this threat for Category I approaches can support any separation, the resulting performance (availability) degradation becomes severe for separations greater than 4 or 5 km.

Finding sufficient space on airport property to site multiple reference receiver antennas while meeting both GBAS and external airport constraints is a challenge. At some airports, such as at Newark, the only possible locations may not meet all GBAS requirements. To account for this, GBAS can adjust several key performance parameters on a site-specific basis. One example is the broadcast $\sigma_{\text{pr_gnd}}$ values that bound the errors of ground-generated pseudorange corrections. Ideally, the same values (as a function of satellite azimuth and elevation) would be broadcast at all sites. However, sites with less-than-ideal multipath performance can adjust this model by broadcasting higher values in bins of azimuth and elevation where reference receiver multipath errors are greater (due to nonideal siting or any other cause). If the errors in certain bins are much higher than ideal, these bins can be *masked*, meaning that corrections are not broadcast for satellites that occupy these bins. *Masking* should only be applied to a small subset of azimuth bins at very low elevation angles (e.g., just above the 5° system elevation mask angle) to avoid significant user performance degradation.

31.3.7 Typical GBAS Errors and Protection Levels

Variants of GBAS that support Category I precision approaches are now fielded at many airports around the world. Most of these are at mid-latitude locations where ionospheric behavior is less severe and better understood, but several sites in equatorial locations (such as Rio de Janeiro, Brazil) have been fielded and are under study.

A good source for recent and near-real-time performance reporting from several GBAS installations is the William J. Hughes FAA Technical Center LAAS website [31.18]. The FAA Technical Center maintains its own LAAS ground station prototype at Atlantic City, NJ as well as providing data from commercial Honeywell SLS-4000 ground stations at Newark, NJ, Houston, TX, Moses Lake, WA, and Rio in Brazil. Using a web interface, it is possible to plot vertical and lateral user errors (based on a static *pseudo-user* antenna not far from the ground station) and vertical and

lateral protection levels at various distances from the ground system centroid and at the approach threshold for each runway supported by the system.

Figure 31.12 illustrates typical GBAS performance in CONUS by plotting vertical accuracy (measured by GBAS versus known from a survey) from the pseudo-user receiver sited near the Newark, NJ (EWR) and Houston, TX (Intercontinental, IAH) airport GBAS sites. These results are conservative because they are dominated by the error in the pseudo-user receiver, which is worse than a typical airborne receiver in that its antenna is on or near the ground and is not an MLA like the ground-station antennas. Even so, typical vertical accuracy is well within 0.5 m with only occasional excursions beyond 0.5 m. The magnitudes and patterns of errors at Newark and Houston are similar, suggesting that GBAS user performance is roughly the same for most mid-latitude locations. This level of nominal accuracy, combined with the *smoothness* of typical error variation (i. e., multipath-induced time correlation over a 150 s approach interval), provides approach guidance to aircraft that is objectively superior to that of ILS [31.63].

Both of these figures cover a week of measurements starting at midnight UT on 3 August 2014, and ending at midnight UT on 10 August 2014. The daily repeti-

tion of the pattern of errors is evident. This is due to multipath repeating every sidereal day (23 h 56 m), when the configuration of GPS satellites returns to the same locations in the sky. Since the position of the pseudo-user antenna relative to reflecting objects is also (mostly) fixed, multipath geometries approximately repeat every sidereal day and produce similar position errors. Significant exceptions to this rule occasionally appear. For example, an error spike just exceeding (negative) 1.5 m appears at Newark around 1600 Coordinated Universal Time (UTC) on 5 August that is not repeated on other days, and the same is true (with a smaller error) at Houston around 1600 UTC on 4 August. These unrepeated error spikes could have several causes but are most likely due to multipath from moving objects (including, possibly, aircraft moving near the pseudo-user antenna).

Figure 31.13 shows the computed GBAS vertical protection levels (VPLs) as estimated by the ground system at each of the runway ends supported by the Newark/EWR and Houston/IAH GBAS systems, respectively. Note that these plots cover only a single 24 h period of repeatable GPS satellite geometries on 3 August 2014. Because VPLs vary mostly with GPS satellite geometry, they will closely repeat from sidereal day to day.

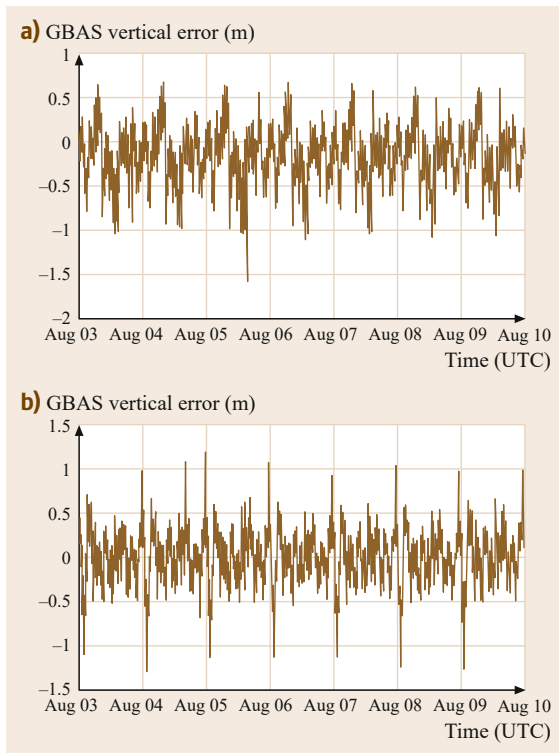


Fig. 31.12a,b GBAS vertical accuracy at Newark (a) and Houston (b) from 3–10 August 2014

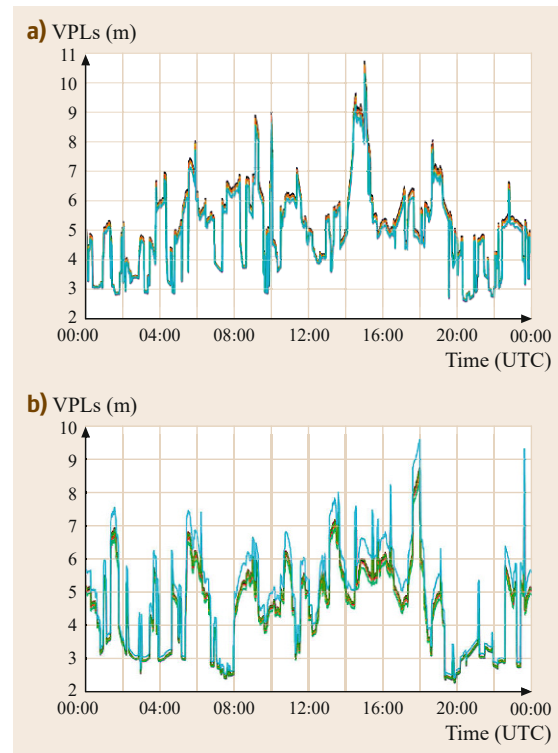


Fig. 31.13a,b GBAS vertical protection levels (VPLs) at Newark (a) and Houston (b) on 3 August 2014

The variation in VPL seen in these figures results partly from changes in the quality of the satellite geometry for positioning as measured by weighted vertical dilution of precision, or **VDOP** (Chap. 1) and partly from the deliberate inflation of σ_{vig} as part of *geometry screening* to protect against worst-case ionospheric gradients (Sect. 31.3.5). VPL values above 7–8 m generally indicate the presence of significant σ_{vig} inflation, which occasionally leads to brief losses of Category I precision approach availability when VPL grows to exceed the 10 m Category I VAL. As shown in Fig. 31.13a, this occurred at Newark/EWR for a brief period around 15:00 UTC for all five runway ends supported by GBAS. Since the degree of σ_{vig} inflation is driven by satellite geometry rather than any particular knowledge of the ionospheric state, this condition will typically persist day after day until the satellite constellation itself changes (e.g., an unhealthy satellite is brought back into service).

31.3.8 Existing GBAS Ground Systems and Airborne Equipment

As noted above, GBAS ground and airborne equipment that supports Category I precision approaches has

been certified and available for several years. The most prominent GBAS ground system on the market is the Honeywell *SmartPath* SLS-4000, which has been fielded at several sites in the US and Europe as well as Rio de Janeiro in Brazil and which generated the data shown in the previous section [31.64]. Other GBAS ground systems include the Thales DGRS 610/615 Reference Station and the NPPF Spektr LCCS-A-2000, which has been fielded at many airports within Russia [31.65].

On the airborne side, the first equipment to be approved for use with GBAS was the Rockwell Collins GLU-925 multimode receiver (MMR), which combines GPS/GBAS (sometimes denoted as *GLS* for *GPS-based landing system*) with ILS and microwave landing system (MLS) receiver components in order to support all sources of precision-approach guidance within a single unit [31.66, 67]. Honeywell (RMA-55B) and Thales (TLS-755) now also offer MMRs that are GPS- and GBAS-capable. Current equipment has only been approved for Category I precision approach, but the developing standards for GAST-D GBAS mentioned earlier provide a pathway to upgrade this equipment to Category II/III approach capability.

31.4 Augmentation via Ranging Signals Pseudolites

31.4.1 Origins and Use in Local-Area DGNSS

Pseudolites are terrestrial devices that transmit GPS or GNSS-like signals to nearby users in order to either enhance the performance of GNSS satellite users or completely replace the need for satellites. One of the first uses of pseudolites was to support the testing of GPS and user equipment functions during the very early days of GPS (the late 1970s), when no or very few satellites were in orbit [31.68]. By the early 1990s, as GPS user equipment became smaller and more portable, small and lightweight pseudolites also became practical and became a key component of early DGNSS systems.

One prototype DGNSS system aimed at providing centimeter-level accuracy (via carrier-phase DGPS) and integrity sufficient to support Category II/III landings was known as the integrity beacon landing system, or **IBLS** [31.3]. The layout of IBLS is shown in Fig. 31.14. Its most unique feature is the use of a pair of pseudolites (*integrity beacons*) sited along each approach path. These pseudolites broadcast weak, carrier-only GPS-like signals (received on antennas mounted on the bottom of the aircraft) that create a *bubble* through which each approaching aircraft flies. The rapid change

in aircraft-to-bubble geometry created by this allows the aircraft to reliably resolve carrier-phase integer ambiguities to both the pseudolites and the GPS satellites. Carrier-phase receiver autonomous integrity monitoring (**RAIM**) at the aircraft is the primary source of integrity monitoring [31.69]. The nearby ground system provides pseudorange and carrier-phase corrections for both satellites and pseudolites but is not responsible for the degree of integrity monitoring described above for GBAS. The quality of pseudorange measurements can be much weaker as well – the primary need for

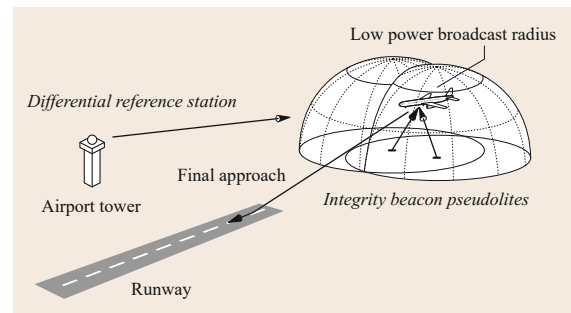


Fig. 31.14 Integrity beacon landing system (IBLS) concept diagram

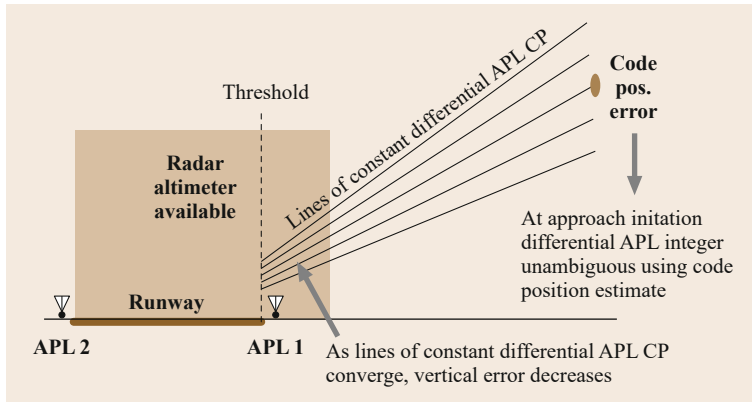


Fig. 31.15 Intrack airport pseudolite (APL) concept diagram (after [31.10])

pseudorange accuracy is to ensure the reliability of the carrier-phase integer ambiguity resolution procedure.

While IBSL worked very well and was demonstrated by multiple aircraft flight tests in the early 1990s [31.3], its requirements for pseudolites along each approach direction (analogous to ILS) and its focus on verifying integrity at the aircraft rather than within the ground system made it unpopular with the FAA and other civil navigation service providers. The IBSL concept was modified to address these concerns in what was known as the *Intrack airport pseudolite (APL) landing system* that was demonstrated at Moffett Field Naval Air Station in the late 1990s. The Intrack airport pseudolite (APL) concept is illustrated in Fig. 31.15. It includes two pseudolites placed at both ends of each runway, but both runway ends are served rather than just one (approaches are supported from either direction). The pseudolites transmit GPS-like pseudorange and carrier-phase signals on unused PRN codes. Approaching aircraft can employ algorithms similar to those in IBSL but do not expect to resolve integer ambiguities. Instead, if *floating* integer values are utilized, the orientation of the two pseudolites along the direction of approach results in improved vertical position accuracy and integrity (where it is most needed), while lateral accuracy and integrity are unaffected. Airborne RAIM is included in this concept, but unlike IBSL, the focus of integrity monitoring is within the ground system that provides differential corrections.

While the Intrack-APL approach requires fewer pseudolites than does IBSL, the need for two pseudolites per runway was seen as excessive for a system that was to serve all runways and to be much easier to site than ILS. The advent of the specialized, multipath-resistant ground system antennas or *MLAs* as described above provided improved pseudorange accuracy and reduced the need for the improvements provided by the Intrack APL layout [31.61, 62]. As a result, the role of

pseudolites in GBAS became that of an additional GPS-like ranging signal broadcast from the ground that was designed to fill in for occasional gaps in GPS satellite coverage. By the early 2000s, even this level of pseudolite augmentation seemed unnecessary, which is why pseudolites are not a feature of today's fielded GBAS systems.

31.4.2 New-Generation Pseudolite Systems for Commercial Applications

In recent years, pseudolite systems designed to more heavily augment (or even replace GNSS) satellites have been developed, particularly for applications where sky visibility is significantly restricted. Open-pit mining is one such application, as equipment operating on the floor of a deep mine is far below the ground's surface and may be unable to see a sufficient number of satellites above the walls of the mine to support accurate positioning.

One pseudolite system developed by Novariant, Inc. to address this scenario is known as *Terralite XPS* and is described in [31.2]. It transmits its own XPS ranging signals in the X-band, between 9.5 and 10 GHz, to avoid interfering with L-band and other commercial transmissions in industrial, scientific and medical (ISM) bands. The combination of XPS signals from multiple pseudolites and at least some L1 and L2 signals from GPS satellites supports three-frequency carrier-phase differential GNSS positioning with accuracies on the order of 10–30 cm. As with the pseudolite systems previously discussed, a differential reference station is needed to provide corrections (raw measurements, in this case) for the L1, L2, and XPS signals received at the reference station to all users. Integrity monitoring is provided within the reference station (along the lines of GBAS) to improve positioning robustness. The number of pseudolites fielded in and around the mine can be adjusted to achieve the needed visibility in the depths of the mine.

In the last several years, pseudolite systems designed to operate mostly or completely independently of GNSS satellites have been introduced. One example has been developed by Locata Corp. of Australia. It is based on individual *LocataLites* that are combined within a network (*LocataNet*) that synchronizes itself wirelessly, which in principle avoids the need for a traditional reference

station. The *LocataLites* transmit on two separate frequencies in the 2.4–2.4835 GHz ISM band, with each frequency being transmitted on two PRN codes. While a *LocataNet* constructed of multiple *LocataLites* can be combined with GPS satellites, it is also capable of operating independently in places where GPS satellites are not available, such as indoors [31.70, 71].

31.5 Outlook

This chapter has summarized the key aspects of ground-based augmentation systems. The GBAS architecture is described in detail as an example of local area differential GNSS (LADGNSS) augmentation systems. Although most LADGNSS systems do not need to meet the integrity and reliability (continuity and availability) requirements of civil aviation as supported by GBAS, the hardware and software components of GBAS are good illustrations of what is possible with LADGNSS.

Fielding pseudolites to augment or even replace GNSS satellites has significant advantages for applications where GNSS-equipped users do not have a full and clear view of the sky. The advent of multiple GNSS constellations makes it possible to receive sufficient

satellite signals even in obstructed locations, but there will likely remain scenarios where pseudolite augmentation remains worthwhile.

Acknowledgments. The author would like to thank the many people in government, academia, and industry who have contributed to the development of GBAS. The author would especially like to note the support of Per Enge and Todd Walter of Stanford, Boris Pervan of the Illinois Institute of Technology, Jiyun Lee of KAIST, John Warburton of the FAA William J. Hughes Technical Center, Tim Murphy and Matt Harris of Boeing, and Mats Brenner and Bruce Johnson of Honeywell. The research support of the US Federal Aviation Administration (FAA) over many years is greatly appreciated.

References

- 31.1 B.W. Parkinson: Introduction and heritage of NAVSTAR, the global positioning system. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996), pp. 3–28, Chap. 1
- 31.2 K.R. Zimmerman, H.S. Cobb, F.N. Bauregger, S. Alban, P.Y. Montgomery, D.G. Lawrence: New GPS augmentation solution: Terralite XPS system for mining applications and initial experience, Proc. ION GNSS, Long Beach (2005) pp. 2775–2788
- 31.3 C.E. Cohen, B.S. Pervan, H.S. Cobb, D.G. Lawrence, J.D. Powell, B.W. Parkinson: Precision landing of aircraft using integrity beacons. In: *Global Positioning System: Theory and Applications*, Vol. 2, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996), pp. 427–459, Chap. 15
- 31.4 F. van Graas, M.S. Braasch: Selective availability. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996), pp. 601–621, Chap. 17
- 31.5 B.W. Parkinson, P. Enge: Differential GPS. In: *Global Positioning System: Theory and Applications*, Vol. 2, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996), pp. 3–50, Chap. 1
- 31.6 W.J. Clinton: *Statement by the President Regarding the United States' Decision to Stop Degrading* *Global Positioning System Accuracy* (White House, Office of the Press Secretary, Washington DC 1 May 2000)
- 31.7 R. Braff: Description of the FAA's local area augmentation system (LAAS), Navigation **44**(4), 411–423 (1997)
- 31.8 GNSS-Based Precision Approach Local Area Augmentation System (LAAS) Signal-in-Space Interface Control Document, DO-246D, 16 Dec. 2008 (RTCA, Washington DC 2008)
- 31.9 Minimum Operational Performance Standards for GPS Local Area Augmentation System Airborne Equipment, DO-253C, 16 Dec. 2008 (RTCA, Washington DC 2008)
- 31.10 B. Pervan, D. Lawrence, K. Gromov, G. Opshaug, J. Christie, P.-Y. Ko, A. Mitelman, S. Pullen, P. Enge, B. Parkinson: Flight test evaluation of an alternative local area augmentation system architecture, Navigation **45**(1), 31–38 (1998)
- 31.11 B. Pervan, F.-C. Chan, D. Gebre-Egziabher, S. Pullen, P. Enge, G. Colby: Performance analysis of carrier-phase DGPS navigation for shipboard landing of aircraft, Navigation **50**(3), 181–192 (2003)
- 31.12 G. Lachapelle, P. Alves: DGPS RTK positioning using a reference network, Proc. ION GPS, Salt Lake City (2000) pp. 1165–1171

- 31.13 G. Crosby, D. Kraus: A ground-based regional augmentation system (GRAS) – The Australian proposal, Proc. ION GPS, Salt Lake City (2000) pp. 713–721
- 31.14 Minimum Operational Performance Standards for GPS Ground-Based Regional Augmentation System Airborne Equipment, DO-310, 13 Mar. 2008 (RTCA, Washington DC 2008)
- 31.15 G. Johnson, C. Oates: USCG NDGPS accuracy and spatial decorrelation assessment, Proc. ION GNSS, Nashville (2012) pp. 3665–3674
- 31.16 S. Pullen, J. Lee: Guidance, navigation, and separation assurance for local-area UAV networks: Putting the pieces together, Proc. ION Pacific PNT, Honolulu (2013) pp. 902–914
- 31.17 IP-921 Product Page (Microhard Systems Inc., Calgary 2014) <http://www.microhardcorp.com/IP921B.php>
- 31.18 Engineering Development Services Group LAAS, US Federal Aviation Administration, William J. Hughes Technical Center, Atlantic City, <http://laas.tc.faa.gov/>
- 31.19 Minimum Aviation System Performance Standards for the Local Area Augmentation System (LAAS), DO-245A, Dec. 9, 2004 (RTCA, Washington DC 2004)
- 31.20 Global Positioning System Wide Area Augmentation System (WAAS) Performance Standard, 1st ed., 31 Oct. 2008 (US Federal Aviation Administration, Washington DC 2008)
- 31.21 T. Murphy, M. Harris, C. Shively, L. Azoulai, M. Brenner: Fault modeling for GBAS airworthiness assessments, Navigation **59**(2), 145–161 (2012)
- 31.22 T. Dautermann, M.L. Felux, A. Grosch: Approach service type D evaluation of the DLR GBAS testbed, GPS Solutions **16**(3), 375–387 (2012)
- 31.23 J. Rife, S. Pullen: Aviation applications. In: *GNSS Applications and Methods*, ed. by S. Gleason, D. Gebre-Egziabher (Artech House, Norwood 2009) pp. 245–267
- 31.24 Specification: Performance Type One Local Area Augmentation System Ground Facility, FAA-E-2937A, 17 Apr. 2002 (US Federal Aviation Administration, Washington DC 2002)
- 31.25 G. Xie, S. Pullen: Integrity design and updated test results for the stanford LAAS integrity monitor testbed, Proc. ION AM, Albuquerque (2001) pp. 681–693
- 31.26 A. Mitelman: Signal Quality Monitoring for GPS Augmentation Systems, Ph. D. Thesis (Stanford Univ., Dept. Aeronautics and Astronautics, Stanford 2004)
- 31.27 G. Wong, R.E. Phelts, T. Walter, P. Enge: Bounding errors caused by nominal GNSS signal deformations, Proc. ION GNSS, Portland (2011) pp. 2657–2664
- 31.28 R.E. Phelts: Multicorrelator Techniques for Robust Mitigation of Threats to GPS Signal Quality, Ph. D. Thesis (Stanford Univ., Dept. Aeronautics and Astronautics, Stanford 2001)
- 31.29 R.E. Phelts, T. Walter, P. Enge: Toward real-time SQM for WAAS: Improved detection techniques, Proc. ION GPS/GNSS, Portland (2003) pp. 2739–2749
- 31.30 F. Liu, M. Brenner, C.Y. Tang: Signal deformation monitoring scheme implemented in a prototype local area augmentation system ground installation, Proc. ION GNSS, Fort Worth (2006) pp. 367–380
- 31.31 FAA LAAS Ground Facility Functions (LAAS KTA Group, Washington DC 1998)
- 31.32 A.J. van Dierendonck: GPS receivers. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996), pp. 329–407, Chap. 8
- 31.33 B. Pervan, D.V. Simili: Code-carrier divergence monitoring for the GPS local area augmentation system, Proc. IEEE/ION PLANS, San Diego (2006) pp. 483–493
- 31.34 B. Pervan, L. Gratton: Orbit ephemeris monitors for local area differential GPS, IEEE Trans. Aerosp. Electron. Syst. **41**(2), 449–460 (2005)
- 31.35 GPS SPS PAN Report #58 (FAA William J. Hughes Technical Ctr., Atlantic City 2007) http://www.nstb.tc.faa.gov/reports/pan58_0707.pdf
- 31.36 H. Tang, S. Pullen, P. Enge, L. Gratton, B. Pervan, M. Brenner, J. Scheitlin, P. Kline: Ephemeris type A fault analysis and mitigation for LAAS, Proc. IEEE/ION PLANS, Indian Wells (2010) pp. 654–666
- 31.37 G. Xie: Optimal On-Airport Monitoring of the Integrity of GPS-Based Landing Systems, Ph. D. Thesis (Stanford Univ., Dept. Aeronautics and Astronautics, Stanford 2004)
- 31.38 Global Positioning System Standard Positioning Service Performance Standard (GPS SPS PS), 4th edn. (US Department of Defense, Washington DC 2008)
- 31.39 S. Pullen, J. Rife, P. Enge: Prior probability model development to support system safety verification in the presence of anomalies, Proc. IEEE/ION PLANS, San Diego (2006) pp. 1127–1136
- 31.40 J. Lee, S. Pullen, P. Enge: Sigma-mean monitoring for the local area augmentation of GPS, IEEE Trans. Aerosp. Electron. Syst. **42**(2), 625–635 (2006)
- 31.41 J. Rife, S. Pullen, B. Pervan: Core overbounding and its implications for LAAS integrity, Proc. ION GNSS, Long Beach (2004) pp. 2810–2821
- 31.42 J. Rife, B. Pervan: Overbounding revisited: Toward a more practical approach for error modeling in safety-critical applications, Proc. ION GNSS, Savannah (2009) pp. 1225–1235
- 31.43 T. Dautermann, C. Mayer, F. Antreich, A. Konovaltsev, B. Belabbas, U. Kalberer: Non-Gaussian error modeling for GBAS integrity assessment, IEEE Trans. Aerosp. Electron. Syst. **48**(1), 693–706 (2012)
- 31.44 J. Rife, R.E. Phelts: Formulation of a time-varying maximum allowable error for ground-based augmentation systems, IEEE Trans. Aerosp. Electron. Syst. **44**(2), 548–560 (2008)
- 31.45 J.R. Clynch, A.A. Parker, R.W. Adler, W.R. Vincent, P. McGill, G. Badger: The hunt for RFI: Unjamming a coast harbor, GPS World **14**(1), 16–23 (2003)
- 31.46 J. Grabowski: Field observations of personal privacy devices, Proc. ION ITM, Newport Beach (2012) pp. 689–741
- 31.47 S. Pullen, G.X. Gao: GNSS jamming in the name of privacy-potential threat to GPS aviation, Inside GNSS **7**(2), 34–43 (2012)
- 31.48 E. Steindl, W. Dunkel: The impact of interference caused by GPS repeaters on GNSS receivers and services, Proc. ENC 2013, Vienna (Austrian Inst. Nav-

- igation, Vienna 2013)
- 31.49 J. Lee, S. Pullen, S. Datta-Barua, P. Enge: Assessment of ionosphere spatial decorrelation for global positioning system-based aircraft landing systems, *AIAA J. Aircraft* **44**(5), 1662–1669 (2007)
 - 31.50 S. Pullen: The use of threat models in aviation safety assurance: Advantages and pitfalls, *Proc. CERGAL 2014*, Dresden (German Inst. Navigation, Bonn 2014)
 - 31.51 S. Pullen, Y.S. Park, P. Enge: Impact and mitigation of ionospheric anomalies on ground-based augmentation of GNSS, *Radio Sci.* **44**(1), RS0A21 (2009)
 - 31.52 S. Datta-Barua, J. Lee, S. Pullen, M. Luo, A. Ene, D. Qiu, G. Zhang, P. Enge: Ionospheric threat parameterization for local area global-positioning-system-based aircraft landing systems, *AIAA J. Aircraft* **47**(4), 1141–1151 (2010)
 - 31.53 M. Kim, Y. Choi, H.-S. Jun, J. Lee: GBAS ionospheric threat model assessment for category I operation in the Korean region, *GPS Solutions* **19**(3), 443–456 (2015)
 - 31.54 C.A. Shively, R. Niles: Safety concepts for mitigation of ionospheric anomaly errors in GBAS, *Proc. ION NTM*, San Diego (2008) pp. 367–381
 - 31.55 S. Ramakrishnan, J. Lee, S. Pullen, P. Enge: Targeted ephemeris decorrelation parameter inflation for improved LAAS availability during severe ionosphere anomalies, *Proc. ION NTM*, San Diego (2008) pp. 354–366
 - 31.56 J. Seo, J. Lee, S. Pullen, P. Enge, S. Close: Targeted parameter inflation within ground-based augmentation systems to minimize anomalous ionospheric impact, *AIAA J. Aircraft* **49**(2), 587–599 (2012)
 - 31.57 T. Murphy, M. Harris: More ionosphere anomaly mitigation considerations for category II/III GBAS, *Proc. ION GNSS*, Fort Worth (2007) pp. 438–452
 - 31.58 S. Khanafseh, S. Pullen, J. Warburton: Carrier phase ionospheric gradient ground monitor for GBAS with experimental validation, *Navigation* **59**(1), 51–60 (2012)
 - 31.59 J. Jing, S. Khanafseh, S. Langel, F.-C. Chan, B. Pervan: Multi-dimensional ionospheric gradient detection for GBAS, *Proc. ION ITM*, San Diego (2013) pp. 121–128
 - 31.60 D. Lamb: Development of local area augmentation system siting criteria, *Proc. ION AM*, Albuquerque (2001) pp. 669–680
 - 31.61 D.B. Thornberg, D.S. Thornberg, M.F. DiBenedetto, M.S. Braasch, F. Graas, C. Bartone: LAAS integrated multipath-limiting antenna, *Navigation* **50**(2), 117–130 (2003)
 - 31.62 A.R. Lopez: Calibration of LAAS reference antennas, *Proc. ION GPS*, Salt Lake City (2001) pp. 1209–1218
 - 31.63 M. Felux, T. Dautermann, H. Becker: GBAS landing system-precision approach guidance after ILS, *Aircraft Eng. Aerosp. Technol.* **85**(5), 382–388 (2013)
 - 31.64 Honeywell: SmartPath Ground-Based Augmentation System (GBAS), <https://aerospace.honeywell.com/products/safety-systems/smart-path>
 - 31.65 N.P.P.F. Spektr: *GBAS activities in Russia, ICAO EUR GBAS Implement. Workshop*, Paris (ICAO, Montréal 2010) pp. 1–53
 - 31.66 C.R. Spitzer, U. Ferrell, T. Ferrell: *Digital Avionics Handbook*, 3rd edn. (CRC, Boca Raton 2014)
 - 31.67 Rockwell Collins: GNLU-9X5M Multi-Mode Receiver, http://www.rockwellcollins.com/Data/Products/Navigation_and_Guidance/Radio_Navigation_and_Landing/GNLU-9X5M_Multi-Mode_Receivers.aspx
 - 31.68 L. Kruczynski: Joint program office test results. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996), pp. 699–715, Chap. 19
 - 31.69 B. Pervan: Navigation Integrity for Aircraft Precision Landing Using the Global Positioning System, Ph. D. Thesis (Stanford Univ., Dept. Aeronautics and Astronautics, Stanford 1996)
 - 31.70 Locata Technology Brief v.8.0 (Locata Corporation, Canberra 2014) <http://www.locata.com/technology/>
 - 31.71 C. Rizos: Locata: A positioning system for indoor and outdoor applications where GNSS does not work, *Proc. APAS 2013*, Canberra (Assoc. Public Authority Surv., Canberra 2013) pp. 73–83

Space Applications

Oliver Montenbruck

Signals transmitted by global navigation satellite system (GNSS) satellites are not confined to the surface of the Earth but can likewise be used for navigation in space. Satellites in low Earth orbits, in particular, benefit from a similar signal strength and experience a full-sky visibility. On the other hand, the harsh space environment, long-term reliability requirements and the high dynamics of the host platform pose specific challenges to the design and operation of space-borne GNSS receivers. Despite these constraints, satellite manufacturers and scientists have early on started to exploit the benefits of GNSS technology. From the first flight of a Global Positioning System (GPS) receiver on Landsat-4, GNSS receivers have evolved into indispensable and ubiquitous tools for navigation and control of space vehicles.

Following a general introduction, the chapter first describes the specific aspects of GNSS signal tracking in space and highlights the technological challenges of space-borne receiver design. Subsequently, the use of GNSS for spacecraft navigation is discussed taking into account both real-time navigation and precise orbit determination. Relevant algorithms and software tools are discussed and the currently achieved performance is presented based on actual missions and flight results. A de-

32.1	Flying High	933
32.1.1	GNSS Tracking in Space.....	934
32.1.2	Spaceborne GPS Receivers	936
32.2	Spacecraft Navigation	938
32.2.1	Trajectory Models.....	939
32.2.2	Real-Time Navigation.....	942
32.2.3	Precise Orbit Determination.....	946
32.3	Formation Flying and Rendezvous	951
32.3.1	Differential Observations and Models ..	952
32.3.2	Estimation Concepts	954
32.3.3	Ambiguity Resolution.....	955
32.3.4	Flight Demonstrations.....	955
32.4	Other Applications	957
32.4.1	Attitude Determination	957
32.4.2	Ballistic Missions	958
32.4.3	GNSS Radio Science.....	959
	References	959

icated section is devoted to the use of space-borne GNSS for relative navigation of formation flying satellites.

The chapter concludes with an outlook on special applications such as spacecraft attitude determination, GNSS tracking of ballistic vehicles as well as GNSS radio science.

32.1 Flying High

Global navigation satellite systems such as GPS, GLONASS, Galileo and BeiDou are primarily designed to offer a worldwide positioning and timing service for ground-based, maritime and airborne users. It is obvious, though, that navigation signals transmitted from the high-altitude GNSS satellites are not just confined to the surface of the Earth and its atmosphere but spill widely into space and can likewise be received by Earth orbiting satellites.

Compared to conventional ground-based tracking systems, the use of GNSS receivers offers a wide range of benefits for spacecraft operations [32.1, 2]. GNSS

tracking can provide an almost continuous coverage and is not limited to short contacts with a ground station. It offers high accuracy at favorable overall system cost and, most notably, enables an increased autonomy by providing navigation information on board a spacecraft rather than generating this information in an offline orbit determination process [32.3].

The potential benefits soon led to first experiments and flight applications of spaceborne GPS receivers. As early as 1982, a time when GPS comprised a mere five active satellites, the flight of the GPSPAC system on board Landsat-4 demonstrated the feasibility of GPS

signal tracking in space and applied such measurements for real-time navigation of the host satellite [32.4]. A decade later, but still well before the buildup of the nominal GPS constellation, the use of a six-channel dual-frequency receiver on board the TOPEX/Poseidon satellite opened the era of GPS-based geodetic-grade orbit determination [32.5, 6]. Extending the use of GPS beyond pure navigation applications, spaceborne radio occultation measurements could first be collected within the GPS/MET experiment on board MicroLab-1 in 1996 [32.7].

The present section discusses the specific constraints of GNSS signal tracking from a spaceborne platform. The visibility and link budgets as well as the signal dynamics experienced in various types of user orbits are first described. Subsequently, the impact of such conditions on the design of spaceborne GNSS receivers is described and an overview of present and upcoming receiver technology is provided.

32.1.1 GNSS Tracking in Space

Most artificial satellites launched into space so far orbit the Earth in a low Earth orbit (LEO) with altitudes of about 200–2000 km and representative periods of 1.5–2 h. Orbits of this type are typically used for remote sensing and surveillance missions but also for various communication constellations such as Globalstar and Iridium. The geostationary Earth orbit (GEO) represent another popular type of orbits that is commonly used for telecommunication and weather satellites. At a distance of roughly 42 000 km from the center of the Earth the orbital period just matches that of the Earth rotation and a satellite appears stationary relative to the surface of the Earth when placed in the equatorial plane. Highly elliptical orbits (HEOs) that stretch from a point close to the Earth up to distances of multiple Earth radii may serve as transfer orbits from LEO to GEO but are also employed for scientific missions exploring the magnetosphere or carrying astronomical telescopes. GNSS satellites themselves are mostly operated in medium altitude Earth orbits (MEOs) with periods near half a day and an orbital radius of about 25 000–30 000 km.

As illustrated in Fig. 32.1, LEO satellites equipped with a GNSS receiver and a zenith-pointing antenna can typically enjoy a GNSS signal coverage that is close to that of a terrestrial user. The received signal strengths are likewise similar, since the altitude of a LEO satellite is generally small compared to the overall distance of the GNSS satellites. It may be noted, though, that the line-of-sight vector extends to larger boresight angles. While terrestrial GPS users remain within a boresight angle of about 14° at all times, a satellite at 1500 km altitude may experience peak boresight angles of close

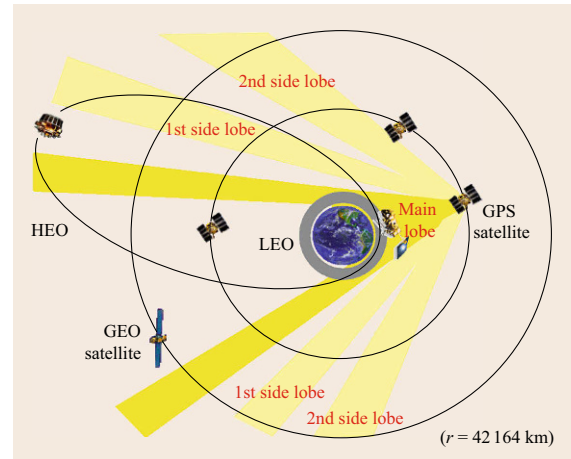


Fig. 32.1 Schematic view of GNSS visibility conditions for LEO, HEO and GEO satellites

to 18° . Even though the transmit antenna gain may be reduced by up to 5 dB in such extreme cases, a LEO satellite always remains within the main lobe of the transmit antenna.

The situation is widely different, though, for HEO and GEO satellites, which even exceed the altitude of common GNSS satellites for most or all of their orbit. Here, GNSS signals can best be received when the transmitting GNSS satellite and the user are on opposite sides of the Earth (Fig. 32.1). For geostationary satellites, the distance from the transmitting GNSS satellite then amounts to almost 70 000 km. This results in large free-space losses and signal levels that are about 10 dB lower than for a terrestrial receiver. In combination with the limited beamwidth of the GNSS transmit antenna it becomes extremely challenging to simultaneously track the minimum of four GNSS satellites that would be required for an independent navigation fix.

While the problem will in part be alleviated by the future availability of multiple GNSS constellations that can be jointly tracked and used for navigation, the tracking of sidelobe signals is often proposed as an alternative. Sidelobes represent local maxima of the antenna gain pattern located at increasingly larger boresight angles. For GPS Block IIA and IIR satellites, first and second sidelobes occur near boresight angles of 30° and 55° but notable differences exist between the various antenna array types on top of strong azimuth variations (Fig. 32.2).

Various successful demonstrations of main and sidelobe signal tracking from spacecraft above the GPS constellation were made between 1997 and 2002 as part of the Falcon Gold experiment [32.9], the Equator-S [32.10] and AO-40 missions [32.11] as well as a nondisclosed GEO project [32.12]. More recently,

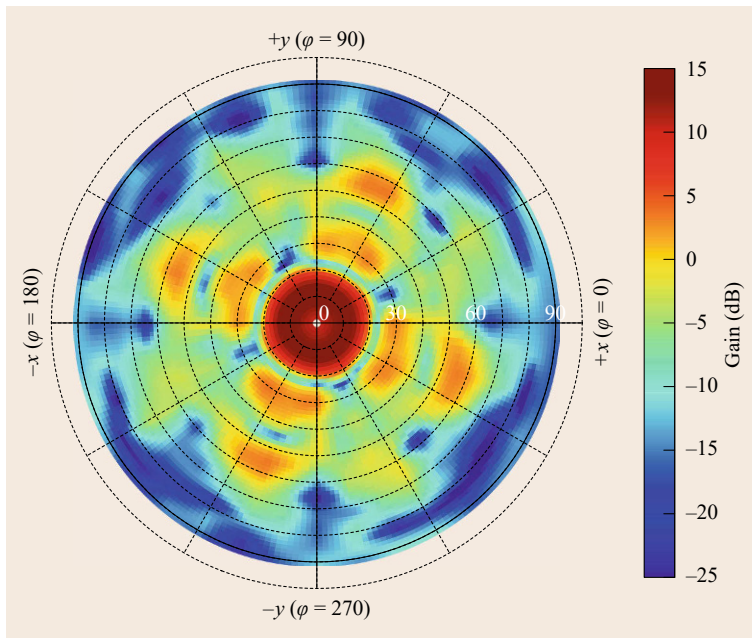


Fig. 32.2 Gain pattern of GPS Block IIR-M transmit antenna on the L1 frequency. The polar plot shows the central main lobe as well as various sidelobes extending up to 60° (after [32.8])

successful sidelobe tracking has been demonstrated with the SGR-GEO receiver onboard the GIOVE-A (Galileo In-Orbit Validation Element) satellite [32.13] and the high-sensitivity navigator receiver on the Magnetosphere Multiscale Mission, MMS [32.14, 15]. It is evident, though, that neither the current GPS nor any other GNSS have been specifically designed with space applications in mind. However, the concept of a new Space Service Volume has been introduced along with the specification of the next generation GPS III satellites [32.16]. Here, minimum received signal levels, signal-in-space range errors and signal availability are addressed for spaceborne users covering altitudes of 3000–36 000 km.

Even though the early tests show that GNSS tracking at high altitudes is indeed feasible, the number of HEO and GEO missions proposing its use for navigation is still very low in view of the limited overall signal availability and performance. The remainder of this chapter is therefore limited to LEO missions where GPS is already widely employed today and new GNSSs that will be supported by the next generation of space receivers.

While link budgets and visibility conditions are generally favorable in such orbits, the tracking conditions are still quite different from those of a terrestrial or airborne user due to the high speed of the host platform. LEO satellites orbit the Earth with a representative velocity of 7.5 km/s and experience a line-of-sight range rate of up to 8.5 km/s when tracking a GNSS satellite (Fig. 32.3). At a wavelength of 0.2 m, a peak Doppler

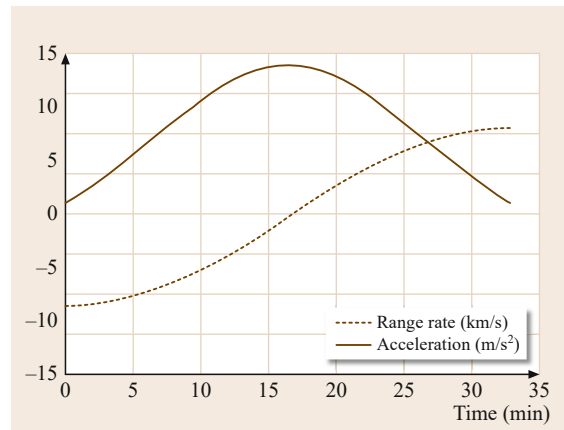


Fig. 32.3 Representative line-of-sight range rate and acceleration for GNSS tracking from a LEO satellite

shift of ± 45 kHz may thus be encountered that is almost ten times higher than for a static receiver on the ground. Likewise, much larger line-of-sight accelerations (close to 15 m/s^2 or 75 Hz/s) are encountered in a circular low Earth orbit. Finally, the visibility period during which a GNSS satellite can continuously be tracked is typically limited to half an hour or less.

Spaceborne GNSS receivers need to cope with these conditions and require some adaptation to accommodate the higher signal dynamics and the rapidly changing set of visible satellites. As an example, particular care must be taken to accurately time-tag all observations. Even a microsecond offset would show

up as an along-track position error of about 1 cm. Furthermore, the tracking loops (Chap. 14) must be of adequate order to avoid steady-state errors due to line-of-sight accelerations. As a rule of thumb, third-order phase-locked loops (PLLs) have to be employed for carrier-tracking, while a carrier-aided first-order delay-locked loop (DLL) may be applied for tracking of the pseudorandom code. Concerning the initial signal search, the large range of possible Doppler results in a pronounced increase of the overall acquisition time compared to terrestrial receivers. When using a traditional sequential search, cold start times of up to 15 min may be required for GPS C/A code receivers in a low Earth orbit and even worse conditions would apply for Galileo E1 open service signals due to the increased code length and the more frequent bit transitions. To assist a rapid signal acquisition, a priori knowledge of the host vehicle position and velocity is therefore commonly employed. This may, for example, be obtained from a numerical or analytical orbit propagator along with a configurable set of orbital elements.

32.1.2 Spaceborne GPS Receivers

Even excluding the problems of high signal dynamics and specific visibility conditions, the design of spaceborne GNSS receivers (Fig. 32.4) poses notable challenges to their manufacturers. As for any kind of spacecraft electronics, a GNSS receiver intended for long-term operations in space needs to fulfill high reliability requirements despite the unique and partly hostile environmental conditions [32.17]. Key aspects that need to be taken care of in the design and premission

verification of space hardware include the robustness against vibration during launch (which may cause physical damage to electrical and mechanical connections), the combined effect of temperature and vacuum (which limits heat dissipation and triggers outgassing of components), and finally space radiation.

The latter is of particular concern for the survivability in space and may affect space electronics in various ways. Ionizing radiation causes a gradual degradation and a continuously increasing power consumption, which ultimately limits the lifetime of the affected components. For satellites in low Earth orbit a total ionization dose (TID) tolerance of 10–20 krad is typically required, but substantially higher limits of up to 100 krad may apply for HEO and GEO missions. Single event upsets (SEUs) and single event latchups (SELs), on the other hand, represent an instantaneous failure of a component caused by the incidence of a high-energy particle. While SEUs (such as bit flips) represent a temporary error that can be overcome by a power cycling or reprogramming, a latchup destroys the affected circuitry on a permanent basis. In complementary metal oxide semiconductor (CMOS) devices, SEUs may arise from parasitic thyristors that are triggered by a short current spike upon incidence of a high-energy particle. Once activated, the thyristor acts as a persistent shortcut and may result in overheating unless the power supply is deactivated immediately [32.18]. Other than TID effects, which may in part be reduced by external shielding, SEUs and SELs require different means of compensation (redundancy, error detection and correction) or protection (rapid fuses, latchup immune semiconductor technology).

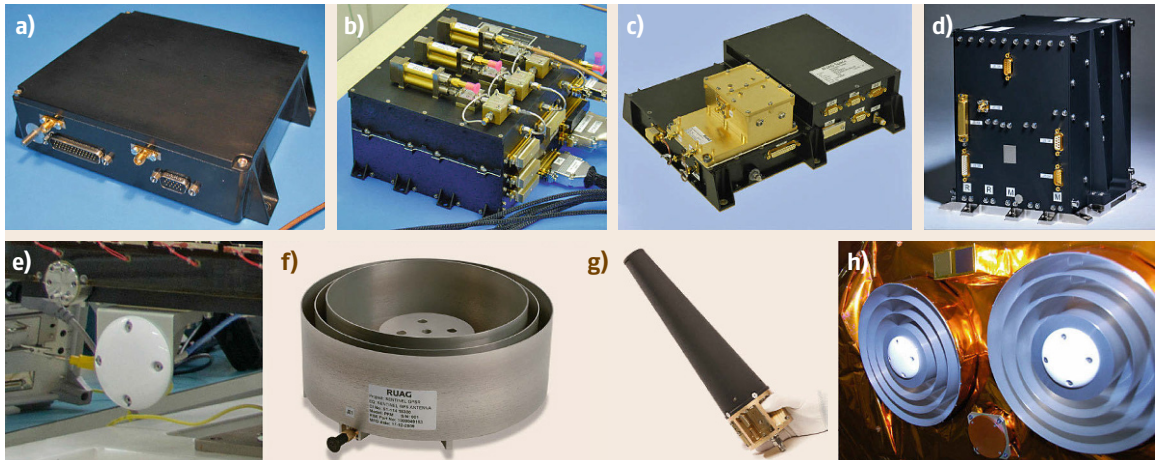


Fig. 32.4a–h Examples of spaceborne GNSS receivers and antennas: (a) SSTL SGR-10 single-frequency GPS receiver, (b) Broadreach Integrated GPS Occultation Receiver (IGOR), (c) RUAG dual-frequency GPS precise orbit determination (POD) receiver for Sentinel, (d) Airbus LION GPS/Galileo receiver; (e) patch, (f) patch excited cup, (g) helix, (h) choke ring antennas. Images not to scale (courtesy of SSTL (a), DLR (b), RUAG (c,f,g), ESA (e) and Airbus DS (d,h))

Table 32.1 Single- and dual-frequency GNSS receivers for space applications

Receiver	Manufacturer (country)	Channels signals	Antennas	Power mass	TID (krad)	Missions
SGR-10	SSTL (UK)	24 GPS L1 C/A	2	5.5 W 1 kg	10	Tsinghua-1, BILSAT, DART
Mosaic GNSS	Airbus DS (D)	8 GPS L1 C/A	1	10 W 4 kg	> 30	SARLupe, TerraSAR-X, Aeolus
TopStar 3000	Thales-Alenia (F)	12-16 GPS L1 C/A	1-4	1.5 W 1.5 kg	> 30	Demeter, Kompsat-2
Viceroy	General Dynamics (US)	12-18 GPS L1 C/A	1-2	7 W 1.1 kg	15	MSTI-3, Seastar, MIR, Orbview
Navigator	NASA/GSFC (US)	12 GPS L1 C/A	1	< 30 W < 11 kg	100	Shuttle HSM-4, MMS
Phoenix	DLR (D)	12 GPS L1 C/A	1	0.9 W 0.1 kg	15	PROBA-2 & -V, PRISMA, TET
GNSS S/W Rcv.	Syrlinks (F)	9 GPS L1 C/A, GAL E1	1	5 W 1 kg	10	Taranis
IGOR	Broadreach Eng. (US)	16 × 3 GPS L1 C/A, L1/L2 P(Y)	4	10 W 4.6 kg	20	COSMIC, TerraSAR-X, TanDEM-X
GPS POD	RUAG (A)	8 × 3 GPS L1 C/A, L1/L2 P(Y)	1	8.5 W 2.8 kg	> 20	SWARM, Sentinel, ICESat-2
Lagrange	Thales-Alenia (I)	12 × 3 GPS L1 C/A, L1/L2 P(Y)	1	30 W 5.2 kg	20	Radarsat-2, COSMO-Skymed, GOCE
TriG	JPL, MOOG Broadreach (US)	24 × 2 GPS/GLO L1/L2, (GAL E1/E5a)	4	55 W 6 kg		Formosat-7/COSMIC-2
LION	Airbus DS (D)	36 GPS L1/L2/L5, GAL E1/E5a	1-4	15 W 6 kg	50	SARah, CSO, Metop-SG
PODRIX	RUAG (A)	18 × 2 GPS L1/L2/L5, GAL E1/E5a	1-4	15 W 3 kg	50	SARah, Sentinel

The various measures taken to ensure a high level of robustness against all aspects of the space environment will typically result in a higher mass and power consumption as well as a conservative performance of spaceborne GNSS receivers in comparison with their terrestrial counterparts. At the same time, the overall system costs are substantially higher due to the small manufacturing volume and the extensive test and validation effort required to guarantee the proper function in space.

For illustration, Table 32.1 summarizes key parameters for a variety of spaceborne GNSS receivers in current use or planned for upcoming missions in the near future. A portfolio of different receivers and antennas is, furthermore, shown in Fig. 32.4. In a trade-off between performance requirements and system cost, single-frequency receivers are commonly preferred for all types of platform navigation and timing applications. Use of dual-frequency GPS receivers, in contrast, is lim-

ited to remote sensing and science missions demanding a highly accurate reconstruction of the spacecraft orbit.

With few exceptions, all receivers available and used in orbit so far have been limited to GPS tracking. Chipsets supporting new signals and constellations are only slowly emerging in parallel with the buildup of the new navigation systems. As an example, the European Space Agency (ESA) has initiated the development of the **AGGA-4** Advanced GPS/GLONASS Application specific circuit (**ASIC**), which offers a total of 36 channels and enables tracking of GPS, GLONASS, Galileo and BeiDou open service signals with binary phase-shift keying (**BPSK**) and binary offset carrier (**BOC**) modulation [32.19]. It supercedes the AGGA-2 correlator chip employed in most European high-end space receivers so far and enables more compact multi-GNSS, multifrequency receiver designs.

The high cost and restricted capability of most spaceborne GNSS receivers has inspired various efforts

to employ commercial off-the-shelf (COTS) technology as an alternative to fully space qualified hardware. Here, advantage can be taken of latest advancements in receiver technology, which is particularly attractive for satellite projects aiming at a high level of miniaturization. Following the pioneering work of Surrey University [32.20] the use of COTS components and GNSS receiver boards has therefore become increasingly popular in low-budget science and technology missions. Even though a dedicated qualification program will always be required to gain adequate confidence in the proper function and survivability of a given device, the use of COTS technology may well represent a favorable trade-off between performance, cost and risks. As an example, a geodetic-grade PolARx2 dual-frequency GPS receiver has been employed for precise orbit determination of the TET-1 technology demon-

stration satellite [32.21] and a Triumph triple-frequency multi-GNSS receiver has been selected for the Atom Clock Ensemble in Space (ACES) experiment [32.22] on board the International Space Station (ISS).

The use of miniaturized aeronautical GNSS receivers has also become of interest for numerous university-class nanosatellites and was first demonstrated successfully within the CanX-2 CubeSat project [32.23]. Besides technical considerations, it must be kept in mind, though, that common export regulations [32.24, 25] for dual-use goods inhibit a free trade of GNSS receivers unless confined to use below a height of 18 km (60 000 ft) and a speed of less than 515 m/s (1000 nm/h). This inhibits a plug-and-play use of standard GNSS receivers on satellites or ballistic vehicles irrespective of signal dynamics or environmental conditions.

32.2 Spacecraft Navigation

GNSS-based orbit determination of spacecraft makes extensive use of well-established concepts and algorithms for terrestrial and airborne navigation. Except in the early days of GPS, when Selective Availability was still in place, an undifferenced processing was generally employed both for real-time and offline processing. All considerations of the GNSS observation and positioning model discussed in Chaps. 19 and 21 are equally valid for spaceborne navigation. Pseudorange-based single-point positioning (SPP) and carrier-phase based precise point positioning (PPP, Chap. 25) techniques can likewise be employed to obtain the instantaneous location of a space vehicle from observations of an onboard GNSS receiver with corresponding levels of accuracy.

As a fundamental difference, though, the free motion of a spacecraft under the dominant gravitational attraction of the Earth is not entirely indeterministic and prone to sudden changes, but highly smooth and predictable. Knowledge of the forces acting on a spacecraft enables prediction of its future motion from the current position and velocity. Other than a purely kinematic SPP/PPP-type positioning, which makes no use of prior information on the spacecraft motion, a dynamical orbit determination exploits this knowledge to reduce the overall number of estimation parameters in the adjustment process. Dynamical models act as a constraint for the estimated position of the host vehicle, which reduces the sensitivity to errors or an unfavorable geometry of the GNSS observations. In addition, they enable a prediction of the spacecraft trajectory through outages and periods of limited GNSS tracking.

While a purely kinematic orbit determination (or point positioning) is of interest for selected applications such as investigations of the Earth's gravity field [32.26, 27], dynamic orbit determination concepts are most widely applied today in view of their obvious benefits for robustness and accuracy. In view of the high precision of GNSS observations (specifically that of carrier-phase measurements) it may, however, be challenging or even impossible to model the spacecraft motion with an equal level of accuracy. The concept of reduced dynamic orbit determination [32.5, 6] was therefore developed early on, which presents a natural compromise between the extremes of a purely kinematic and a purely dynamic processing (Fig. 32.5). Here, additional force model parameters (known as empirical or stochastic accelerations) are introduced on top of the deterministic trajectory model and estimated along with other orbit and measurement model parameters. These empirical parameters can be constrained to the expected uncertainties of the a priori force model and allow the adjusted trajectory to better adapt to the true motion as sensed by the GNSS measurements.

Depending on the specific application, orbit models of varying complexity and detail, different estimation concepts as well as different strategies for the incorporation of empirical parameters are employed in the orbit determination process. Real-time onboard navigation and ground-based precise orbit determination constitute two representative examples that will serve to illustrate typical mission needs and practical solutions. Even though the boundaries between low-latency and high-accuracy applications are gradually removed,

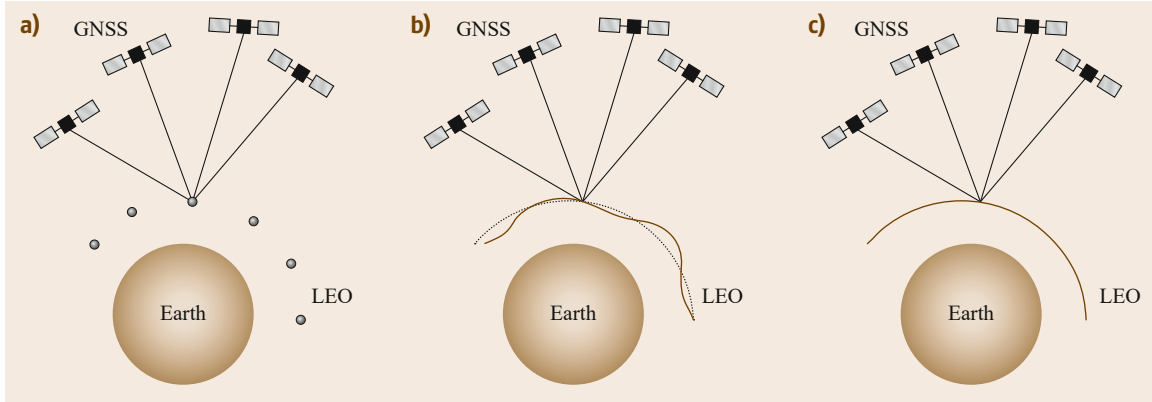


Fig. 32.5a–c GNSS-based precise orbit determination concepts for LEO satellites: (a) kinematic, (b) reduced dynamic, and (c) dynamic orbit determination

the two cases are well suited to highlight the relevant algorithms and concepts. Before going into a detailed discussion of these aspects, the modeling of spacecraft trajectories is reviewed that forms an integral part of any orbit determination process.

32.2.1 Trajectory Models

In its most general form, the orbital motion of a satellite as a function of time t is described by a second-order differential equation

$$\frac{d^2 \mathbf{r}}{dt^2} = \mathbf{a}(t, \mathbf{r}, \mathbf{v}, \mathbf{p}) \quad (32.1)$$

relating the change in position \mathbf{r} to the acceleration \mathbf{a} , which in turn depends on time, position, velocity \mathbf{v} , and additional parameters \mathbf{p} . Equivalently, the equation of motion may be formulated as a first-order differential equation

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(t, \mathbf{y}) = \begin{pmatrix} \mathbf{v} \\ \mathbf{a} \end{pmatrix} \quad (32.2)$$

for the position-velocity (or state) vector $\mathbf{y} = (\mathbf{r}^\top, \mathbf{v}^\top)^\top$. With given initial conditions $\mathbf{y}(t_0) = \mathbf{y}_0$ and known accelerations \mathbf{a} , the state vector at any other time can then be determined from the analytical or, more commonly, numerical solution of this differential equation.

The equation of motion is complemented by the variational equations

$$\frac{d}{dt}(\Phi, \mathbf{S}) = \begin{pmatrix} \mathbf{I} & \mathbf{I} \\ \frac{\partial \mathbf{a}}{\partial \mathbf{r}} & \frac{\partial \mathbf{a}}{\partial \mathbf{v}} \end{pmatrix} (\Phi, \mathbf{S}) + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \mathbf{a}}{\partial \mathbf{p}} \end{pmatrix} \quad (32.3)$$

for the state transition matrix

$$\Phi(t, t_0) = \frac{\partial \mathbf{y}(t)}{\partial \mathbf{y}(t_0)}, \quad (32.4)$$

which describes the dependence of the instantaneous state vector on the initial state, as well as the sensitivity matrix

$$\mathbf{S}(t) = \frac{\partial \mathbf{y}(t)}{\partial \mathbf{p}}, \quad (32.5)$$

which describes the dependence of the state vector on the force model parameters [32.28]. The partial derivatives provided by these matrices are required within the orbit determination process to establish the set of initial conditions and model parameters that best represents a given set of observations.

A variety of numerical integration methods for initial value problems are available in the open literature [32.28–30] and any of them can, in principle, be used for solving the equation of motion and the variational equations of a satellite orbit. However, there is hardly a single best method serving all applications needs. A key criterion for the selection of a suitable integration method is its order n , which characterizes the growth of errors in a single integration step. For an n th-order method the local truncation error is of order $\mathcal{O}(h^{n+1})$ in stepsize h and the computed solution is such of similar accuracy as an n th-order Taylor approximation.

Single-step methods such as Runge–Kutta and extrapolation methods evaluate the derivative \mathbf{f} in the equation of motion (32.1) at selected points within the interval $[t_i, t_i + h]$ to construct an approximation for the state $\mathbf{y}(t_i + h)$ from a given state $\mathbf{y}(t_i)$. Depending on the desired order an increasing number of function evaluations is required, which notably raises the computational effort of high-order methods. However, the increased effort is usually well compensated by the larger stepsizes that can be used. High-order methods are therefore preferable to cover large intervals

with a small total integration error. For stepsize control, approximations of different orders may be used to estimate the local integration error in a given step and to adjust the stepsize in accord with given accuracy requirements. Common implementations of these concepts include the Runge–Kutta–Fehlberg [32.31] methods as well as those of *Dormand and Prince* [32.29, 32].

Multistep methods compute a predicted value

$$\mathbf{y}(t_{i+1}) = \mathbf{y}(t_i) + \int_{t_i}^{t_i+h} \mathbf{p}_f(t) dt \quad (32.6)$$

of the state vector at time $t_{i+1} = t_i + h$ from a polynomial approximation $\mathbf{p}_f(t)$ of the function $\mathbf{f}(t, \mathbf{y})$ at past epochs t_i, t_{i-1}, \dots . At the expense of storing information across multiple epochs, this concept ensures a high efficiency and facilitates the generation of high-order methods. As an added advantage, multistep methods enable a straightforward interpolation of the solution for dense output irrespective of the actual integration stepsize. Common classes of multistep methods include the Adams–Bashforth–Moulton method as well as Stoermer–Cowell and Gauss–Jackson methods. The latter types are particularly efficient for integrating the second-order formulation (32.1) of the equation of motion.

Multistep methods are most easily formulated when working with a fixed stepsize, but become increasingly complex when stepsize adjustments are desired. Also, a special starting scheme is required to establish the tableau of past function values upon initialization or restart of the integrator. Both aspects are readily covered by advanced implementations such as the variable-order and variable-stepsize Adams–Bashforth–Moulton method DDEABM of *Shampine and Gordon* [32.33] or the variable-step double-integration multistep integrator of *Berry* [32.34], which combines the efficiency of traditional multistep methods with the flexibility commonly attributed to single-step methods.

The choice of an integrator for orbit determination is largely tied to the type of estimation method used for the parameter adjustment as well as the interval between observations. Within a batch least-squares estimation the trajectory over the entire data arc is described by a single epoch state vector. Large stepsizes can be used to cover the entire data arc if the observations are sparse enough or if interpolation can be used to obtain the state at individual measurement epochs. Multistep methods appear best suited in this case and have been applied in various precise orbit determination packages such as GEODYN [32.35], NAPEOS [32.36], GIPSY-OASIS [32.37], or GHOST [32.38]. As an alternative, the Bernese software [32.39] makes use of the

collocation method [32.30], which favorably combines the benefits of single- and multistep methods by constructing a Taylor approximation of the solution over a desired time step from the given initial conditions.

Real-time navigation in contrast will typically make use of low-order integration methods, since observations are processed in short intervals and since each measurement update in an extended Kalman filter mandates a restart of the trajectory integration [32.40]. This does not allow the exploitation of the large stepsizes supported by high-order methods and reduces their efficiency in this application. The well-known fourth-order Runge–Kutta (RK4) method provides a good compromise between accuracy and efficiency in this case and lends itself as a general purpose method for trajectory propagation in real-time navigation systems. As suggested in [32.29], the basic RK4 method can also be combined with the concept of Richardson extrapolation to obtain a fifth-order method requiring less function evaluations than a native fifth-order Runge–Kutta method. More significantly, however, the intermediate values computed in this process enable construction of an interpolating Hermite polynomial of consistent order [32.40]. This is particularly useful if high-rate orbit information is required for attitude and orbit control purposes or for geocoding of sensor data onboard the satellite [32.41].

Irrespective of the particular method chosen for the equation of motion, the quality of the dynamical model determines the accuracy of the orbit prediction over a given timespan or data arc. Again, a high-fidelity model is usually adapted in offline orbit determination, while simplicity may be favored in real-time navigation systems with limited computational resources or lacking access to auxiliary parameters such as Earth orientation angles or solar flux.

An overview of different accelerations acting on a satellite in low Earth orbit is given in Table 32.2 [32.28, 42, 43]. Similar to GNSS satellite orbits (Chap. 3), the motion of a LEO satellite is governed by the gravitational attraction of the Earth, even though the asphericity of the Earth gravity field has a much more pronounced effect at low altitudes. Generally speaking, the acceleration is described as the gradient $\mathbf{a}_{\text{grav}} = \nabla V$ of the gravitational potential V , which is itself represented by a spherical harmonics expansion

$$V = \frac{GM_{\oplus}}{r} \sum_{n=0}^{\infty} \sum_{m=0}^n \frac{R_{\oplus}^n}{r^n} \bar{P}_{nm}(\sin \phi) \times [\bar{C}_{nm} \cos(m\lambda) + \bar{S}_{nm} \sin(m\lambda)] \quad (32.7)$$

in terms of geocentric distance r , latitude ϕ and longitude λ as well as the gravitational coefficient (GM_{\oplus}) of

Table 32.2 Representative values of accelerations acting on a LEO satellite

Perturbation	Acceleration
Central gravity term	8.5 m/s ²
Earth oblateness ($C_{2,0}$)	15 mm/s ²
Aspherical gravity field	
10 × 10 versus $C_{2,0}$	0.2 mm/s ²
40 × 40 versus 10 × 10	30 μm/s ²
100 × 100 versus 40 × 40	3 μm/s ²
Air Drag	0.1–10 μm/s ²
Moon	1 μm/s ²
Sun	0.5 μm/s ²
Solid Earth tides	0.5 μm/s ²
Solar radiation pressure	50 nm/s ²
Ocean tides	50 nm/s ²
General relativity	20 nm/s ²

the Earth and the normalized gravity field coefficients ($\bar{C}_{nm}, \bar{S}_{nm}$). The normalized Legendre polynomials \bar{P}_{nm} and the trigonometric functions required for the evaluation of the gravitational potential and its gradient are best obtained through dedicated recurrence relations [32.44]. These offer a high level of computational efficiency and ensure the numerical stability required for use with models of high degree and order.

Gravity models in use today for satellite orbit prediction are themselves derived from geodetic satellite missions such as CHAMP, GRACE, GOCE and LA-GEOS (see [32.45–47] and references therein). While some of these models are also augmented by surface gravity measurements to provide the highest possible resolution, the use of satellite-only models is generally fully adequate for satellite orbit modeling. The required degree and order depends on the envisaged modeling accuracy and may vary from 10 × 10 for simple onboard application to 150 × 150 for geodetic space missions with utmost accuracy requirements.

Aside from the Earth itself, a satellite is attracted by the Sun (☉) and Moon (☾), to mention just the dominant solar system bodies. In a geocentric reference frame the associated third-body perturbation

$$\mathbf{a}_{\odot, \odot} = \sum_{i=\odot, \odot} GM_i \left(\frac{\mathbf{s}_i - \mathbf{r}}{\|\mathbf{s}_i - \mathbf{r}\|^3} - \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|^3} \right) \quad (32.8)$$

is determined by the difference of the respective accelerations exerted on the satellite and the Earth itself. Here GM and s denote the product of the gravitational constant and the perturbing body's mass as well as its geocentric position. The latter may be obtained

from precomputed solar system ephemerides or analytical series expansions [32.28]. Given the proximity of a LEO satellite to the center of the Earth in comparison with the Sun-Moon distance, the two terms within the bracket in (32.8) are of almost similar magnitude and cancel except for a small tidal acceleration that grows almost linearly with the satellites distance from the Earth.

Among the nongravitational accelerations, atmospheric drag is often the most pronounced perturbation on a LEO satellite, particularly when considering altitudes of 350–500 km. The perturbing acceleration is always directed opposite to the satellite's velocity \mathbf{v}_{rel} relative to the atmosphere. This results in a continued loss of orbital energy and a change of the along-track position that grows essentially quadratically in time. Even a small drag acceleration can thus result in substantial orbital perturbations. Aside from the surface-to-mass ratio (A/m), the acceleration

$$\mathbf{a}_{\text{Drag}} = -\frac{1}{2} C_D \frac{A}{m} \rho \|\mathbf{v}_{\text{rel}}\| \mathbf{v}_{\text{rel}} \quad (32.9)$$

depends on the atmospheric density ρ , which may vary by several orders of magnitude within the relevant altitude range. Here, the drag coefficient C_D accounts for the dependence of the drag acceleration on the body shape. It can, in principle, be obtained from computational fluid dynamics (CFD) calculations for a known satellite structure [32.48, 49]. In the absence of such models, approximate values of 2.0–2.3 are commonly adopted as a starting point and refined drag coefficients are adjusted within the orbit determination.

In parallel with an increasing number of space missions, continuous effort has been made since the late 1960s to develop suitable physical and numerical models for the atmospheric density and its variation with time, location and other input parameters. Popular models used for satellite orbit modeling include the Jacchia models (Jacchia-70 and later refinements), the *mass spectrometer* and *incoherent scatter* models (MSIS-86, NRLMSIS-00 [32.50]) and the *drag temperature models* (DTM-2009 [32.51], DTM-2012). The individual models offer a typical accuracy of 10–30% [32.52]. Atmospheric density prediction thus constitutes a major source of uncertainty in the modeling of low-altitude satellite orbits.

Independent of orbital altitude, satellites are subject to a second type of nongravitational accelerations caused by the momentum transfer of absorbed and reflected photons. Similar to atmospheric drag, the resulting acceleration depends on the surface-to-mass ratio and is thus most pronounced for satellites with large solar panels. At a mean distance of one Astronomical Unit (1 AU ≈ 149.6 km) the solar flux amounts

to roughly 1367 Wm^{-2} and the Sun exerts a pressure of

$$P_{\odot} \approx 4.56 \cdot 10^{-6} \text{ Nm}^{-2} \quad (32.10)$$

on a surface that absorbs all incident radiation.

In the most simple form, solar radiation pressure (SRP) can be described by a so-called cannonball model, where the perturbing acceleration

$$a_{\text{SRP}} = -\eta C_R \frac{A}{m} P_{\odot} \left(\frac{1 \text{ AU}}{\|r_{\odot}\|} \right)^2 \frac{r_{\odot}}{\|r_{\odot}\|} \quad (32.11)$$

is always directed opposite to the Sun direction and scales with the inverse square of the distance $\|r_{\odot}\|$ from the Sun. Since satellites may move within the shadow of the Earth (or Moon) for at least part of their orbit, the factor η is used to model the overall illumination. It varies between 1 in full Sun light and 0 in total eclipse. The solar radiation pressure coefficient

$$C_R = 1 + \varepsilon \quad (32.12)$$

depends on the reflectivity ε and takes into account that the impulse transfer is twice as high for a fully reflecting surface as compared to full absorption. Typical materials exhibit reflectivities ranging from 0.2 (solar panels) to 0.9 (coated mylar foil) and an average C_R value of 1.3–1.5 is commonly applied with the cannonball model. For a more refined modeling, individual surface properties and orientations relative to the Sun direction need to be handled in the form of a box-wing model or even a full-featured ray-tracing computation [32.53]. Due to remaining uncertainties in the actual surface properties, it is common practice, though, to adjust at least a single scaling factor (e.g., C_R) of the radiation pressure model as a free parameter within the orbit determination process.

Aside from the major constituents of the force model (Earth gravity, lunisolar gravity, drag, and radiation pressure) introduced above, a multitude of smaller gravitational and nongravitational forces need to be taken into account for high-precision modeling of LEO satellite trajectories [32.28, 42, 43]. These include solid-Earth, pole and ocean tides [32.54], post-Newtonian corrections to the equation of motion [32.30], as well as Earth albedo and thermal radiation effects [32.53]. A detailed discussion of these perturbations and their modeling is beyond the scope of this text and interested readers are referred to the aforementioned text books and articles.

Irrespective of the adopted modeling level and complexity, it frequently turns out that the real-world dynamics are not fully matched by the equation of motion. This is particularly obvious when working with high-precision observations such as GNSS and satellite laser

ranging (SLR) measurements. It is therefore common practice to complement the deterministic a priori force model with empirical accelerations. These accelerations can then be adjusted within the orbit determination such as to best match the observations and the modeled trajectory under suitable constraints on the expected value and magnitude of the empirical parameters.

To facilitate their interpretation, empirical accelerations are typically parameterized in an orbital frame aligned with the radial (e_R), along-track (e_T) and cross-track (e_N) direction

$$a_{\text{emp}} = a_R e_R + a_T e_T + a_N e_N. \quad (32.13)$$

In the most simple form, the individual components a_i ($i = R, T, N$) are treated as constants over a certain time of applicability. Alternatively, additional *once-per-rev* terms may be considered to account for perturbations that exhibit a harmonic variation with the orbital period. The resulting accelerations can then be described as

$$a_i = a_{i,0} + a_{i,c} \cos u + a_{i,s} \sin u, \quad (32.14)$$

where u denotes the argument of latitude.

32.2.2 Real-Time Navigation

Use of GPS (or, more generally, GNSS) receivers onboard LEO satellites has become increasingly popular, since it offers position, velocity and timing information onboard a spacecraft and can therefore contribute to an increased autonomy. Common applications of GNSS-based orbit information include attitude control, geocoding of payload data, autonomous instrument and spacecraft operations as well as orbit control:

- Earth observation (EO) missions commonly require alignment of their instruments (e.g. cameras, altimeters, or light detection and ranging (LIDAR) sensors) with the nadir direction and the ground track. While star cameras provide highly accurate measurements of the spacecraft orientation, the resulting information is naturally referred to a celestial reference frame. Knowledge of the instantaneous position and velocity is thus required to refer the measured attitude to an orbital frame aligned with the nadir and flight direction [32.41]. Onboard position and velocity knowledge with accuracies of about 10 m and 1–10 cm/s is commonly required in EO mission specifications for attitude control support.
- Geocoding refers to the association of spacecraft instrument data (such as pixels within a camera image) with the corresponding geographic location.

Traditionally, it is performed on the ground after determining the spacecraft position (and attitude) with adequate accuracy. Onboard geocoding of payload data (or at least onboard orbit determination) reduces the necessary data processing within the mission control center and supports a direct downlink of preprocessed image products to users in remote locations. Since local terrain models that would be required for a rigorous geocoding are mostly substituted by approximate geoid models in onboard applications, orbit information at the 10 m level is generally compatible with the overall accuracy requirements.

- Onboard position information can further be used to perform autonomous instrument and spacecraft operations. Instead of time-tagged activation, a synthetic aperture radar (SAR) or other sensor may be activated once a required orbital position is reached [32.55]. In this way, the limitations of ground-based mission planning and the dependence on predicted orbit information can be overcome. Similar considerations apply for the operations of the satellite itself. As an example, the spacecraft may autonomously reorient itself and activate its transmitters for performing a data dump once it enters the visibility range of a ground station. Depending on the specific application, onboard orbit information with an accuracy at the 10–1000 m level will be required.
- Last, but not least, orbit information provided by a GNSS sensor can be used to autonomously control the orbital motion (such as the ground track or the equator crossing time) of a remote sensing satellite and to compensate orbit changes caused by atmospheric drag or other natural orbital perturbations [32.56]. A 10–100 m position knowledge is generally adequate for orbit control of individual satellites or a loose constellation. However, the maneuver planning cannot be based on instantaneous position and velocity measurements alone but requires a proper averaging and monitoring of orbital elements variations over timescales of multiple orbits or days.

The above discussion demonstrates that the standard positioning service (SPS) or open service (OS) performance of common GNSSs is fully adequate for the majority of onboard applications related to spacecraft and instrument operations. Even though the desired accuracy can, in principle, be delivered by a standalone GNSS receiver in low Earth orbit, a complementary navigation filter is usually foreseen to ensure proper robustness and continuity of the navigation information. This filter employs a dynamical

model to reduce the inherent measurement noise, to provide orbit information between consecutive measurement epochs and to predict the orbit across periods of limited GNSS tracking. A reduced number of visible GNSS satellites may, for example, be encountered near the poles or during nonzenith pointing orientation of a spacecraft.

Depending on the overall system and redundancy concept, a navigation filter may reside inside the satellite's onboard processor or the GNSS receiver itself. While the former approach offers a great flexibility in the choice of receivers and facilitates handover between prime and backup units, a navigation system inside the GNSS receiver is often advantageous in terms of interfaces and the transferred amount of data. Filtered navigation solutions (of varying accuracy) are already provided by a variety of GNSS receivers (e.g., MosaicGNSS and Topstar3000, Lagrange; Table 32.1) designed specifically for spacecraft operations and platform support.

Generically speaking, a real-time navigation system employs an orbit propagator to predict the satellite orbit and the associated covariance between consecutive GNSS observations. These observations are then processed in a Kalman filter that blends the predicted state with the new measurement information. With each filter update, a new estimate of the current spacecraft position is thus obtained.

Even though the use of alternative filter concepts has occasionally been suggested for GNSS-based onboard navigation, the use of an extended Kalman filter (Sect. 22.5) is most common and fully adequate for the purpose. For typical time steps (of 1–30 s), proper initial conditions and a reasonably accurate force model, a real-time navigation filter operates close to linearity. As such, filter concepts for highly nonlinear problems (e.g., the unscented or sigma-point *Kalman* filter, UKF [32.57]) do not offer relevant advantages that justify the increased computational effort within this type of application.

In the context of satellite orbit determination, the filter state $\mathbf{x} = (\mathbf{y}^T \mathbf{p}^T \mathbf{q}^T)^T$ is composed of the position-velocity vector \mathbf{y} as well as other parameters that need to be adjusted within the estimation. These comprise force model parameters \mathbf{p} (such as drag and radiation pressure coefficients or empirical accelerations) as well as measurements model parameters \mathbf{q} (such as clock offsets and biases).

As discussed in Sect. 22.5, the extended Kalman filter is a recursive estimation scheme, which is continuously repeated as new observations are processed. Given the estimated filter state $\hat{\mathbf{x}}_{i-1|i-1}$ and its covariance $\mathbf{P}_{i-1|i-1}$ at epoch t_{i-1} , the corresponding values at the following epoch t_i are obtained through combination of

a time-update step and a subsequent measurement-update step.

Based on the given position, velocity, and force model parameters at time t_{i-1} , the trajectory is first propagated by numerical integration of the equation of motion (32.2) to obtain the spacecraft state vector \mathbf{y}_i at the subsequent measurement epoch t_i . The remaining filter parameters (\mathbf{p} , \mathbf{q}) are likewise propagated to the new epoch using the corresponding stochastic process models. Empirical accelerations, for example, may be treated as exponentially correlated random variables [32.58], in which case the predicted value

$$\mathbf{a}_{i|i-1} = \mathbf{e}^{\frac{t_i - t_{i-1}}{\tau}} \mathbf{a}_{i-1|i-1} \quad (32.15)$$

results from an exponential damping in accord with the correlation timescale τ . For bias parameters, in contrast, constancy can be assumed and the propagated parameter matches the value at time t_{i-1} .

Complementary to the state propagation, the state transition matrix $\Phi_{i|i-1} = \partial \mathbf{x}_i / \partial \mathbf{x}_{i-1}$ for the combined filter state is formed. It comprises the state transition and sensitivity matrices of the position-velocity vector (which follow from the numerical integration of the variational equations (32.3)) as well as (less computationally intensive) blocks for the remaining estimation parameters. Overall, the time-update step of the extended Kalman filter yields the predicted state vector $\mathbf{x}_{i|i-1}$ along with the corresponding covariance matrix

$$\mathbf{P}_{i|i-1} = \Phi_{i|i-1} \mathbf{P}_{i-1|i-1} \Phi_{i|i-1}^\top + \mathbf{Q}_i, \quad (32.16)$$

where \mathbf{Q}_i denotes the covariance increase due to process noise in the dynamical model.

Based on the difference between the actual measurements \mathbf{z}_i and the modeled observations $\mathbf{g}(\mathbf{x}_{i|i-1})$ an improved estimate

$$\hat{\mathbf{x}}_{i|i} = \hat{\mathbf{x}}_{i|i-1} + \mathbf{K}_i (\mathbf{z}_i - \mathbf{g}(\hat{\mathbf{x}}_{i|i-1})), \quad (32.17)$$

of the state at time t_i is obtained within the measurement-update step. The Kalman gain

$$\mathbf{K}_i = \mathbf{P}_{i|i-1} \mathbf{G}_i^\top (\mathbf{W}_i^{-1} + \mathbf{G}_i \mathbf{P}_{i|i-1} \mathbf{G}_i^\top)^{-1}, \quad (32.18)$$

which provides a linear mapping of the innovation $\mathbf{z} - \mathbf{g}$ to the state domain depends on the measurement covariance \mathbf{W}_i^{-1} , the Jacobian $\mathbf{G} = \partial \mathbf{g}(\mathbf{x}) / \partial \mathbf{x}$ of the observation model as well as the predicted state covariance $\mathbf{P}_{i|i-1}$. It is furthermore used to obtain the postmeasurement-update covariance

$$\mathbf{P}_{i|i} = (\mathbf{I} - \mathbf{K}_i \mathbf{G}_i) \mathbf{P}_{i|i-1} \quad (32.19)$$

of the estimated state, where \mathbf{I} denotes the identity matrix. Together, $\mathbf{x}_{i|i}$ and $\mathbf{P}_{i|i}$ provide the necessary information for the subsequent cycle of the extended Kalman filter.

Concerning the practical use of (32.19), it is noted that symmetry and positive semidefiniteness of the resulting matrix may not be guaranteed in numerical computations with limited precision. This may result in a hard-to-identify source of filter divergence. More robust formulations of the filter equations (such as the UD-factorization [32.59] with upper diagonal factor \mathbf{U} and diagonal factor \mathbf{D}) have therefore been developed and are generally preferred in actual flight software.

The specific concepts and algorithms employed in a real-time onboard navigation system depend largely on the desired accuracy but also on considerations of processor load as well as software complexity, reliability and verification. In the most simple case, kinematic position fixes of a GNSS receiver are processed as observations within a Kalman filter estimating only the instantaneous position and velocity. The approach is well suited for use inside an onboard processor and can be employed in combination with arbitrary GNSS receivers. It requires a minimalistic measurement model and decouples the navigation filter from the intrinsic details of GNSS-based positioning. In particular, no clock-offset parameter needs to be estimated in the filter and no explicit knowledge of the GNSS satellite positions is required for the processing. Sample implementations of a GNSS-position filter are described in [32.28] and [32.41], where representative accuracies of 5 m have been achieved with a single-frequency GPS receiver. Even though the filter yields a smooth and predictable trajectory, it can only partly compensate for errors in the employed position measurements, which are affected by uncompensated ionospheric path delays and the limited precision of the broadcast ephemeris errors. Also, the simple navigation filter is unable to process observations when less than four GNSS satellites are tracked and the receiver is no longer able to provide a kinematic position solution.

Advanced real-time navigation systems will therefore process raw pseudorange and, optionally, carrier-phase observations within the measurement-update step. For pseudorange-only processing, the observation model is essentially the same as that of a standalone GNSS receiver. It comprises the modeling of the light-time corrected geometric distance between the receiver and the GNSS satellite based on broadcast ephemerides as well as receiver and GNSS satellite clock offsets (including relativistic clock corrections) and relevant group delays. Tropospheric corrections are obviously of no concern at orbital altitudes but ionospheric path delays need to be considered through a model of the

individual path delays unless a ionosphere-free linear combination of dual-frequency observations is employed. The *Klobuchar* model [32.60] used in terrestrial receivers is of limited value for use in low Earth orbit, since the peak electron density is typically at or even below the altitude of the receiver. As an alternative, the *Lear* model [32.61] may be employed, which describes the slant delay at elevation E as the product $I = I_0 m(E)$ of the vertical path delay I_0 and a dedicated mapping function

$$m(E) = \frac{2.037}{\sqrt{\sin^2 E + 0.076} + \sin E}. \quad (32.20)$$

The vertical delay itself is rarely known in advance and may vary rapidly along the orbit. It should therefore be incorporated into the filter state and estimated together with other dynamical and measurement model parameters.

Pseudorange-based Kalman filters are commonly adopted for navigation systems inside a spaceborne GNSS receiver. They offer a reasonable compromise between robustness, accuracy and software complexity. For highest accuracy, however, the processing of carrier-phase observations is indispensable. The use of phase measurements requires the incorporation of bias parameters that need to be estimated in the filter state and (re)initialized, whenever a new satellite is allocated to a tracking channel or whenever a cycle slip is encountered. Outlier identification and rejection in a spaceborne carrier-phase navigation filter is often more difficult than in terrestrial applications (among others, due to rapid position changes and variations in the ionospheric path delays, but also due to limited computational resources) and often considered a potential risk for the overall reliability and robustness. Only limited flight experience is therefore available at present.

Following [32.62], the accuracy of a dual-frequency GPS navigation filter can be improved from roughly 1 m to 0.5 m (3-D root mean square [RMS] position error) when incorporating carrier-phase measurements in addition to code observations. The adopted filter design utilizes a fidelity force model including Earth gravity (up to degree and order 70) and lunisolar perturbations as well as drag and solar radiation pressure. Remaining force model deficiencies are compensated by empirical accelerations that are adjusted within the filter. Overall, the filter state

$$\mathbf{x} = (\mathbf{r}^\top \mathbf{v}^\top C_D C_R \mathbf{a}_{\text{emp}}^\top cdt \mathbf{b}^\top)^\top \quad (32.21)$$

incorporates the host satellite position \mathbf{r} and velocity \mathbf{v} , the drag and radiation pressure coefficients (C_D , C_R),

empirical accelerations \mathbf{a}_{emp} , the receiver clock offset cdt and a vector of biases \mathbf{b} representing the carrier-phase ambiguities of all tracking channels. For a representative 12-channel GNSS receiver, a 24-dimensional estimation state is thus obtained.

Evidently, the factor-of-two performance gain achieved by the incorporation of carrier-phase observation in addition to pseudoranges does not reflect the inherent precision of the respective measurements. In fact, the achievable real-time navigation performance is mostly limited by the quality of broadcast ephemerides, which exhibit representative signal-in-space range errors (**SISRE**) of 0.3 m (GPS Block IIF) to 1.2 m (GPS Block IIa) [32.63]. While the adjustment of bias states in the navigation filter can partly help to compensate the impact of slowly varying broadcast orbit errors [32.62], the broadcast clock quality remains a limiting factor for the performance of GNSS real-time navigation systems. For comparison, an onboard orbit determination accuracy of 0.1 m has successfully been demonstrated in the Doppler orbitography and radiopositioning integrated by satellite (**DORIS**) immediate orbit onboard determination (**DIODE**) system of the JASON-2 satellite, which employs Doppler measurements from ground-based radio beacons [32.64]. It can be shown, though, that a similar accuracy is readily achievable in GNSS-based navigation systems when using real-time clock corrections [32.62, 65]. These may, for example, be provided through dedicated communication links as in NASA's tracking and data relay satellite system (**TDRSS**) augmentation service for satellites (**TASS** [32.66]) or via dedicated GNSS services such as the Quasi-Zenith Satellite System (**QZSS**) centimeter level augmentation service [32.67] and the proposed Galileo commercial service.

Aside from dual-frequency pseudorange and carrier-phase observations, the filter concept described above can also be applied with a ionosphere-free combination of single-frequency observations. The **GRAPHIC** (group and phase ionospheric correction) concept was first proposed by *Yunck* [32.68] and has since then been employed for a variety of space missions with single-frequency GPS receivers. It makes use of the fact that ionospheric path delays affect code and phase measurements with opposite sign and cancel (to first order) when forming the arithmetic average

$$o_{\text{GPH}(p,\varphi)} = \frac{1}{2}(p + \varphi) \quad (32.22)$$

of the measured pseudorange (p) and the corresponding carrier-phase measurement (φ). Due to the unknown carrier-phase ambiguity, the **GRAPHIC** observation is a biased measurement but exhibits only half the

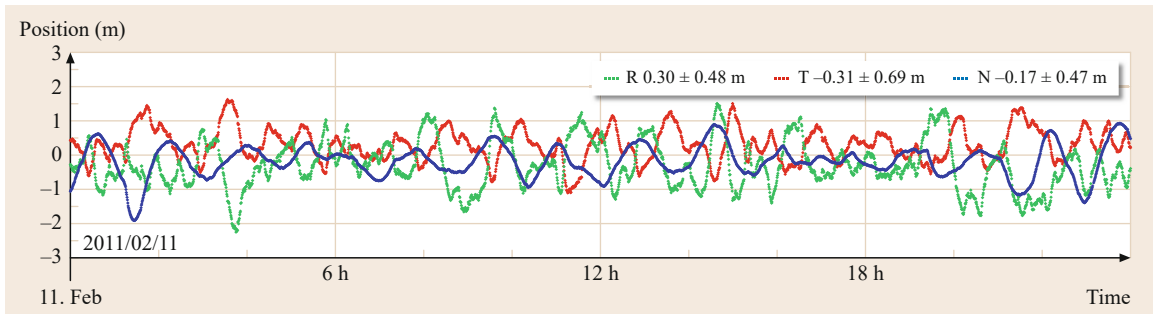


Fig. 32.6 Real-time navigation performance of the Phoenix-XNS navigation system on board the PROBA-2 satellite in radial (R), along-track (T) and cross-track (N) direction

noise level of the pseudorange measurement. Similar to the ionosphere-free carrier-phase combination, the GRAPHIC measurement can be processed in a Kalman filter that adjusts the relevant ambiguities along with other estimation parameters.

As an example of an actual flight application, the performance of the Phoenix extended navigation system (XNS) on board the PROBA-2 spacecraft of the European Space Agency (ESA) is shown [32.69]. The XNS software is integrated into a miniaturized 12-channel GPS receiver and processes GRAPHIC observations based on L1 C/A code and phase measurements. As illustrated in Fig. 32.6 a steady-state accuracy of about 1.1 m (3-D root mean square (RMS) position error relative to a ground based precise orbit determination result) has been demonstrated in actual flight tests and an even better performance of about 0.7 m could be achieved in postflight analyses through use of current Earth orientation parameters and a refined filter tuning.

32.2.3 Precise Orbit Determination

Precise orbit determination (POD) commonly refers to estimation of a satellite's position and velocity with the highest possible accuracy. Other than real-time on-board navigation, POD is performed on the ground and can thus make use of the most elaborate processing schemes, powerful computer infrastructure, and the best available auxiliary data. Traditionally, latency is of less concern than quality. Delivery times up to several weeks are often accepted for the best and final POD solutions.

In much the same way as tracking systems and modeling techniques have been refined over the past decades, the notion of precision has likewise changed by several orders of magnitude. GNSS observations (as well as other techniques such as SLR or DORIS) enable orbit determination accuracies at the subdecimeter level on a routine basis and a 1 cm 3-D RMS position accuracy is within reach of the most advanced POD approaches. Along with that, processing times have

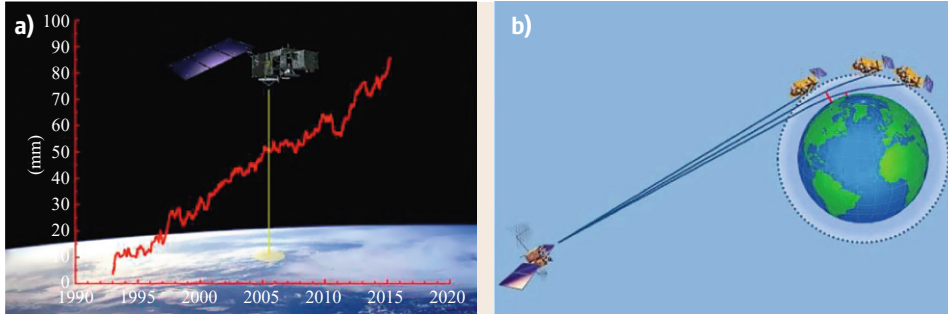
decreased dramatically and high-quality orbit determination results can today be delivered with latencies of several hours and less.

The quest for ever-improved POD accuracies (Table 32.3) is mainly driven by the overall objectives and the specific instrumentation of scientific space missions:

- Synthetic aperture radar (SAR) missions commonly require position accuracies close to the instrument resolution (approx. 1 m) for SAR image generation [32.72]. These values mainly drive the need for near-real-time orbit determination. A (sub)decimeter accuracy, in contrast, is required for the less time-critical interferometric processing of images collected during repeated passes over the same scene as well as high-precision radar ranging applications [32.73].
- Altimeter missions determine the mean sea level from the known satellite position and measurements of its height above the ocean surface (Fig. 32.7a). For climatology, the sea level must be monitored with centimeter to millimeter level accuracy over decades, which requires a corresponding quality of the orbit determination [32.70]. Uncertainties of less than 1–2 cm in radial direction are therefore commonly required for altimeter missions [32.74].
- GNSS radio occultation (RO) missions (Chap. 38) monitor the troposphere by measuring the signal delay and the associated bending of the signal path when a GNSS signal traverses the Earth's atmosphere (Fig. 32.7b). The bending angle can be inferred from the difference of the observed Doppler shift and that of a fictitious straight-line signal. Knowledge of the along-track velocity with an uncertainty of less than 0.05–0.2 mm/s is therefore required [32.75]. This is roughly equivalent to a 5–20 cm position knowledge.
- For gravity missions, such as GOCE and GRACE, varying accuracy requirements arise from mission

Table 32.3 POD requirements of selected science missions (R,T,N: radial component, along-track, cross-track component; 3-D: total error)

Mission	Type	Near-real-time	Final
GOCE	Gravity	0.5 m 3-D	-
TerraSAR-X	search and rescue (SAR)	1 m 3-D	10 cm 3-D
Sentinel-1	SAR	10 cm 3-D	5 cm 3-D
Sentinel-2	Optical	1 m 3-D	-
Sentinel-3	Altimetry	8 cm R	2 cm R
Jason-1/2	Altimetry	-	1.5 cm R
Metop-A	radio occultation (RO)	0.1 mm/s T	

**Fig. 32.7a,b** Earth observations requiring high-precision orbit determination to achieve their science goals. **(a)** ESA's Sentinel-3 satellite measuring the sea surface height with a high-precision altimeter (courtesy of ESA). The overlay shows the variation of the global mean sea level as deduced from altimeter measurements of the Topex and Jason satellites. (After [32.70, 71]). **(b)** GPS radio occultation measurements with the Metop satellite. (Courtesy of EUMETSAT, US Gov)

and instrument operations (e.g., a subnanosecond time synchronization of two spacecraft within a formation). Furthermore, purely kinematic positioning solutions with accuracies of a few centimeters are commonly requested to support an independent recovery of low-degree and order components of the gravity field from the observed spacecraft motion.

In order to meet the most stringent accuracy requirements, GNSS-based precise orbit determination combines sophisticated dynamic models with the concepts of carrier-phase-based precise point positioning (PPP; Chap. 25). Even though double-differencing with respect to ground-based reference stations has been employed in early POD concepts (namely before the abandoning of selective availability in GPS), undifferenced approaches are almost exclusively employed today.

Except for tropospheric path delays, which are of no concern at orbital altitudes, the GNSS observation model for LEO orbit determination considers the same terms and corrections as introduced in Chap. 19 for terrestrial and airborne applications. This includes the light-time modeling, relativistic clock and range corrections, phase center offsets and variations of the GNSS satellite and the receiving antenna, group and phase delays as well as phase wind-up effects. Since the

dynamical model describes the motion of the center-of-gravity (COG) of a space vehicle, proper knowledge of its attitude and the location of the antenna reference point relative to the COG is a vital prerequisite for precise orbit determination. Given the fact that LEO satellites move by more than 7 m within a millisecond, a proper time-tagging of all measurements and a careful distinction between receiver time and GNSS system time (or the respective fundamental time scale used in the equation of motion) is, furthermore, required.

Most software packages for GNSS processing and LEO orbit determination (such as GEODYN [32.35], NAPEOS [32.36], or GHOST [32.38]) make use of a weighted least squares estimation scheme to adjust dynamical orbit parameters and GNSS observation model parameters within the orbit determination. While details vary among the different software implementations and configurations, a representative estimation parameter vector

$$\mathbf{X} = (\mathbf{T}^\top \mathbf{Y}^\top \mathbf{B}^\top)^\top \quad (32.23)$$

will comprise a vector

$$\mathbf{T} = (cdt_1, \dots, cdt_{n_T})^\top \quad (32.24)$$

of receiver clock offset parameters for a total of n_T measurement epochs, the dynamical model parameters

$$\mathbf{Y} = (\mathbf{r}^\top, \mathbf{v}^\top, C_D, C_R, \mathbf{a}_{\text{emp},1}^\top, \dots, \mathbf{a}_{\text{emp},n_A}^\top)^\top, \quad (32.25)$$

which include the initial position and velocity vector (or equivalently a vector of six initial orbital elements), the drag and radiation pressure parameters, as well as a vector of empirical accelerations for a total of n_A intervals, and, finally, a vector of (float-valued) carrier-phase ambiguities

$$\mathbf{B} = (b_1, \dots, b_{n_B})^\top \quad (32.26)$$

for a total of n_B passes of continuous carrier-phase tracking.

Considering a 30 s sampling rate, approximately 3000 clock offset parameters have to be adjusted for a one-day data arc when processing observations of a single constellation. In a multi-GNSS orbit determination constellation-specific clock offsets or intersystem biases need to be adjusted at each epoch, which results in an associated increase in the total number of estimation parameters [32.76]. The number of empirical acceleration parameters employed in a LEO orbit determination depends on the quality of the dynamical model and may range from a small number of one-per-rev parameters in a highly dynamical POD configuration to a large set of piecewise constant accelerations in a reduced-dynamic approach. Considering, for example, one set of accelerations in radial, along-track and cross-track direction per 10 min interval (i. e., roughly 1/10th of the orbital period) about 500 acceleration parameters need to be adjusted in a 24 h orbit determination arc. The number of ambiguity parameters, finally, depends on the total number of satellites in a constellation and the number of orbital revolutions (assuming that each GNSS satellite is tracked once per orbit). Some 500 parameters per day are typically required for GPS-only tracking unless the dataset is affected by frequent cycle slips. For future multi-GNSS POD applications, the number of estimation parameters will again increase in accord with the amount of tracked satellites and constellations.

Given the large number of estimation parameters, the direct solution of the full normal equations using Gauss elimination, LU-decomposition, or Cholesky factorization would become fairly computation intensive. It may be noted, though, that the normal equations matrix (Fig. 32.8) is dominated by a sparse diagonal matrix related to the clock-offset parameter vector \mathbf{T} . This matrix can easily be inverted, which enables a pre-elimination of the respective parameters and yields

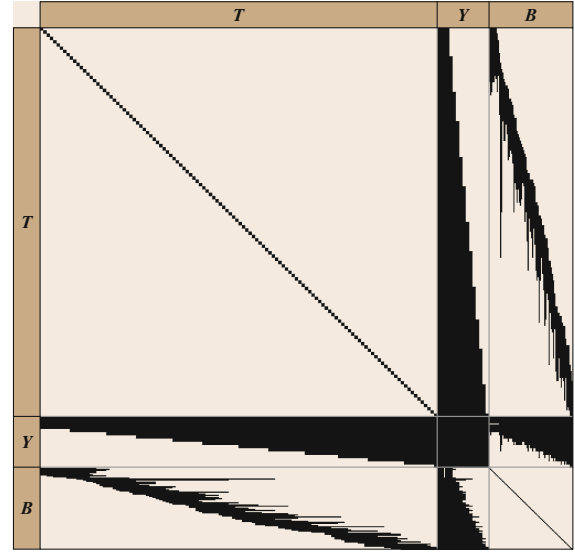


Fig. 32.8 Representative structure of normal equations for GNSS-based precise orbit determination of a LEO satellite. *Light brown areas* denote cells with zero values

a reduced set of normal equations for the remaining dynamical (\mathbf{Y}) and ambiguity (\mathbf{B}) parameters and enables an efficient solution of the overall normal equations [32.77].

The estimation of piecewise-constant empirical accelerations is illustrated in Fig. 32.9, which shows the along-track accelerations required in addition to the a priori force model to best fit the GPS observations of the TerraSAR-X satellite during a one-day data arc in 2007. The acceleration is assumed to be constant within each 10 min interval. The observation residuals exhibit a standard deviation of about 7 mm for the ionosphere-free L1/L2 carrier-phase combination, which corresponds to a noise level of about 2 mm for the individual single-frequency phase observations. To achieve this goodness of fit, systematic phase pattern variations of the employed choke ring antenna (Fig. 32.4) have been corrected through a phase center variation (PCV) map (Fig. 32.10) derived from an inflight calibration.

Given the fact that GNSS-based precise orbit determination is itself one of the most accurate (and certainly the most widely used) techniques for measuring the position of a LEO satellite in space, it is inherently difficult to validate the performance that can actually be achieved. Other than in terrestrial PPP applications, where the analysis of time series for static monitoring stations enables a direct assessment of the positioning performance, there is hardly any truth standard for the motion of a space vehicle. Nevertheless, a variety of internal and external quality measures can be utilized

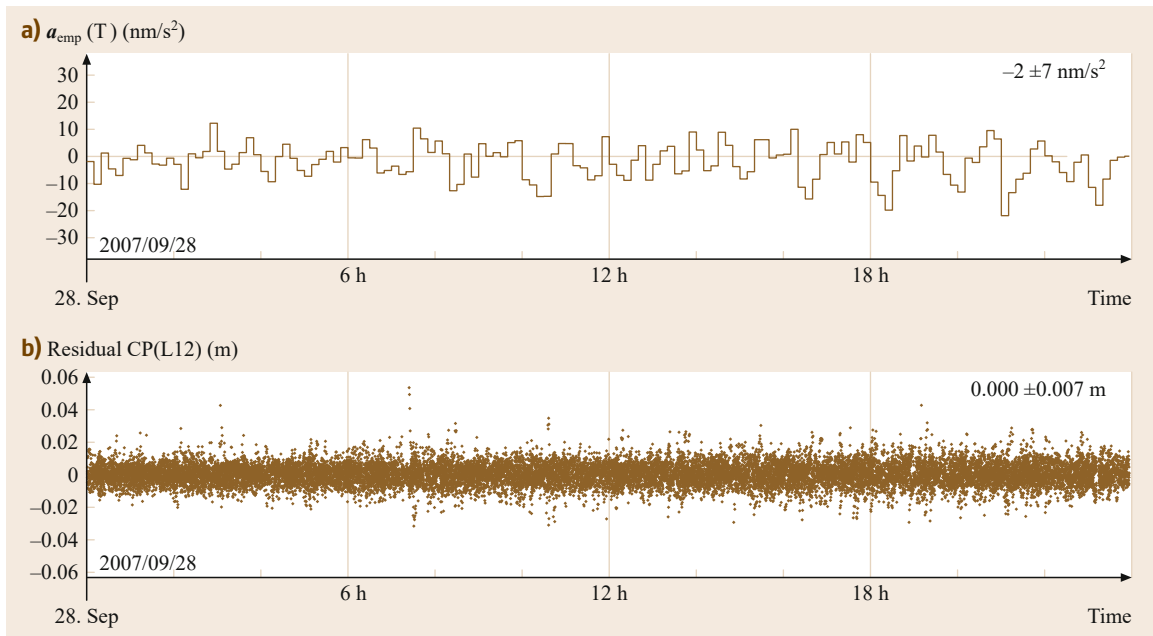


Fig. 32.9a,b Example of empirical accelerations compensating the difference between true and modeled dynamics for the TerraSAR-X satellite (a) and resulting carrier-phase residuals (b)

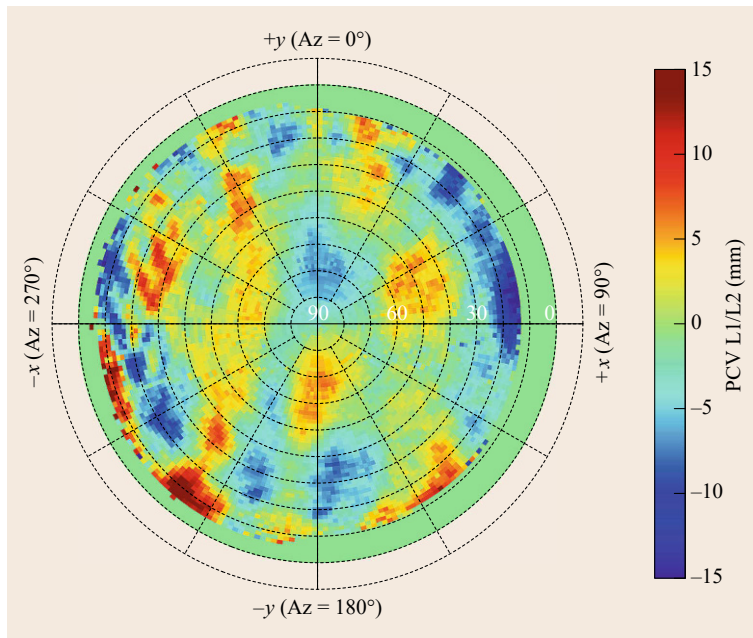


Fig. 32.10 Phase center variations of TerraSAR-X POD antenna

to characterize the accuracy of a precise orbit determination solution. These provide good confidence that the specified performance (Table 32.3) can in fact be reached and even superseded.

In the absence of external comparison methods the achievable orbit determination quality can be assessed

through various types of (self-)consistency checks. Within interagency comparisons the sensitivity of the resulting solution to the employed algorithms, processing standards and auxiliary data products can be evaluated using different software tools for a common dataset. Even though all solutions are based on the same

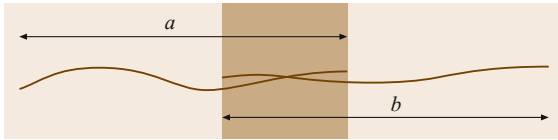


Fig. 32.11 Overlap comparisons provide a measure of the self-consistency of two different orbit solutions (a, b) within the common part of the data arc



Fig. 32.12 Satellite laser ranging station of the Zimmerwald observatory (courtesy of Astronomisches Institut der Universität Bern)

GNSS observations, the resulting orbit solutions will hardly be identical due to different processing schemes and the resulting differences provide a fair measure of the overall accuracy. As an example, a 5 cm 3-D RMS consistency has been demonstrated for Metop-A orbit determination results obtained by five institutions using different software packages and different levels of dynamic versus reduced dynamic processing [32.78].

As an alternative, overlap comparisons are frequently used for quality control and performance assessments. Orbit determination solutions using different data arcs will usually differ when comparing epochs in the common overlap region (Fig. 32.11). Such differences are most pronounced near the start and tail of each POD solution and can be used as an overall quality measure. Overlap comparisons provide a good performance indicator in a highly dynamic orbit determination, but tend to be less representative of the true accuracy in a reduced dynamic approach. Here the solutions are strongly driven by the observations and tend to align to each other in the overlap region. For a more significant POD assessment, other validation methods are therefore desired, which are based on independent measurement systems of comparable accuracy or precision.

Among the external validation techniques, satellite laser ranging (SLR) represents probably the most universal and widely used method for assessing the quality of GNSS-based precise orbit determination solutions.

SLR measures the turnaround time of short laser pulses and provides a highly precise and unambiguous measure of the distance between a satellite and the SLR station. The use of SLR stations (Fig. 32.12) for scientific space missions is coordinated by the International Laser Ranging Service (ILRS, [32.79]), which supervises the operation of 40–50 stations around the world and ensures a priority-based tracking of individual satellites.

As a fully passive system, a laser ranging reflector can easily be accommodated on the host satellite and is widely used in geodetic space missions. Even though SLR observations can also be used for orbit determination as a standalone measurement type or in combination with other observations, the available number of measurements is generally substantially lower than that of GNSS due to the sparse ground network and the much higher operational effort. For satellites equipped with geodetic-grade GNSS receivers it is therefore of primary interest as an independent validation tool. An example of SLR residuals relative to GNSS-based orbit solutions is shown in Fig. 32.13 for the GOCE gravity mission. Reduced-dynamic orbits for this mission have been computed by the Astronomical Institute of the University of Bern (AIUB) from dual-frequency GPS observations of the Lagrange receiver [32.80]. The SLR residuals exhibit a standard deviation of 1.5 cm and a mm-level bias, which demonstrates an excellent consistency of both tracking techniques and a GPS POD accuracy of about 2–3 cm.

Similar POD performances have been reported for the Jason-1 and -2 satellites. In addition to GPS and SLR, these satellites are also equipped with a DORIS (Doppler orbitography and radiopositioning integrated by satellite) receiver, that provides Doppler measurements relative to a global ground station network. DORIS observations can be used for independent or combined POD solutions and an excellent consistency is obtained for all three techniques, which demonstrates a 3-D accuracy of 2–3 cm [32.81]. For the radial component of the Jason orbits, which is of primary interest for altimetry, an accuracy of 7–9 mm has been determined in the comparison of GPS-based POD solutions with SLR observations, SLR/DORIS orbits and altimeter crossovers [32.82].

While the accuracies discussed above are certainly remarkable and largely fulfill present-day requirements, continuous effort is made to further reduce the error level of GNSS-based LEO orbit determination solutions. A joint adjustment of GNSS satellite orbits and multiple LEO satellites in a common orbit determination process with full ambiguity fixing is often considered as a means for improving both the overall accuracy and the reference frame tie, but has not been conducted so far in view of the excessive com-

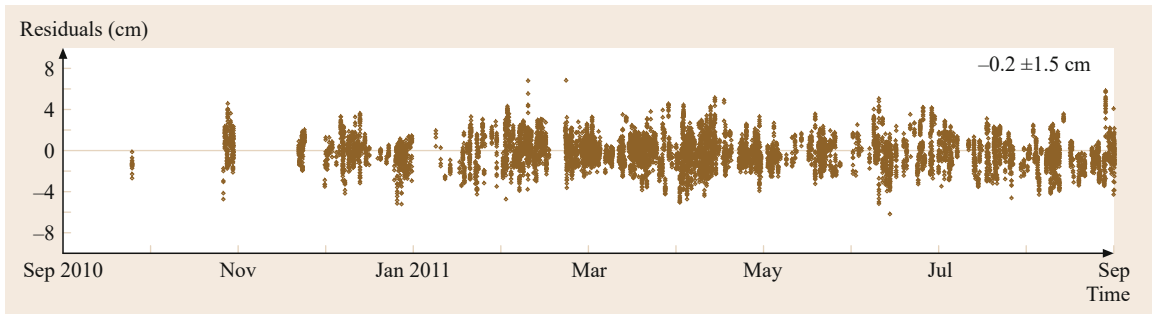


Fig. 32.13 SLR residuals of Astronomical Institute of the University of Bern (AIUB) precise orbit determination solutions for the GOCE satellite (September 2010–August 2011)

putational effort. However, single-receiver ambiguity fixing concepts that are likewise employed in modern PPP concepts (Sect. 25.3.4), have been shown to provide a favorable alternative. Various approaches have been proposed that make use of dedicated GNSS clock products (so-called integer clocks [32.83]) or phase bias products [32.84] to convey the underlying network information but avoid the explicit use of ground-station observations within the POD process. Single-receiver ambiguity fixing has been shown to improve both the absolute and the relative accuracy of LEO satellite orbits in sample applications to the GRACE and Jason satellites [32.83, 84]. It also appears as a particularly promising technique for precise orbit determination of loose constellations with multiple satellites at baselines of hundreds to thousands of kilometers.

Complementary to accuracy, the timeliness of POD solutions is of growing interest in many remote sensing missions to enable a rapid release of science products and a quick-look data analysis. For GNSS-based precise

orbit determination, the availability of sufficiently accurate low-latency GNSS orbit and clock data has long been a bottleneck in the generation of rapid POD solutions. Various institutions have therefore established independent ground networks and services (such as JPL's global differential GPS (GDGPS, [32.66]) system or ESA's ground support network (GSN, [32.85])) to meet the reliability and latency needs of dedicated space missions. More recently, similar products have also become publicly available as part of the International GNSS Service (IGS) real-time service (RTS, [32.86]). Even though the IGS does not offer a formal service guarantee, its products benefit from a large and highly redundant network infrastructure and are provided free of charge to its users. As shown in various studies [32.87, 88], advanced (near-)real-time GNSS orbit and clock products can nowadays be used to generate near-real-time LEO POD solutions that exhibit only a minor performance degradation compared to their final counterparts.

32.3 Formation Flying and Rendezvous

As discussed in the preceding sections, spaceborne GNSS receivers and PPP-style processing techniques have found a wide range of applications for the navigation of individual spacecraft. Depending on the level of effort and the algorithms involved, accuracies ranging from a few meters in real-time single-point positioning to several centimeters in precise orbit determination may be achieved. It has long been realized, though, that differential GNSS (DGNSS) would be a powerful tool for relative navigation of multiple spacecraft. Similar to terrestrial applications, various forms of common errors (such as ionospheric path delays as well as GNSS orbit and clock errors) are partly or fully eliminated when working with differential measurements over short baselines. Also,

carrier-phase ambiguities can more easily be resolved among nearby spacecraft than in space-to-ground baselines, which require a detailed consideration of atmospheric path delays. Carrier-phase-based differential navigation of orbiting spacecraft can therefore offer an order-of-magnitude increase in (relative) accuracy compared to absolute GNSS positioning techniques (Fig. 32.14). Furthermore, high-precision relative navigation can also be achieved in real-time applications due the reduced dependence on GNSS ephemeris information.

Applications that can directly benefit from GNSS-based relative navigation include spacecraft formation flying as well as orbital rendezvous with cooperative targets. While formation flying aims at a long-term op-

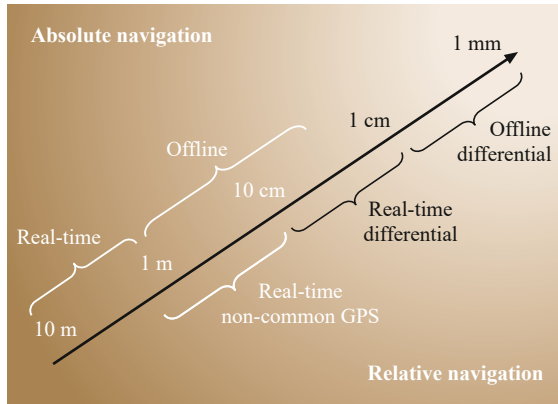


Fig. 32.14 Comparison of absolute and relative positioning accuracies achievable in spaceborne GNSS navigation (after [32.89])

eration of two or more spacecraft in close proximity to achieve an advanced mission goal (such as gravimetry or interferometry [32.90]), rendezvous is usually a short-term activity but covers a wider range of relative distances. Accurate knowledge of the relative position and velocity is a prerequisite for both types of missions to properly control the motion of all participating spacecraft in accordance with the overall mission concept and the safety requirements [32.91]. Furthermore, precise relative orbit determination is often essential for achieving the primary science goals of a formation flying mission. In all cases GNSS recommends itself as an attractive navigation sensor in view of its performance, cost and onboard availability. Relative GPS has indeed helped to realize a variety of ambitious space missions, some of which will be highlighted in Sect. 32.3.4.

In accordance with earlier considerations, the subsequent discussion focuses on relative navigation in low Earth orbit, which offers the best GNSS visibility

and coverage conditions. Even though formation flying and rendezvous are also discussed for high-latitude and even deep-space missions, such missions will typically make use of other types of radiometric or optical navigation sensors.

32.3.1 Differential Observations and Models

Spaceborne relative navigation using GNSS builds on well established differential GNSS processing concepts and combines these with knowledge of the (relative) orbital motion of the involved spacecraft [32.89].

Either single- or double-difference observations are most widely employed for spaceborne relative navigation (Fig. 32.15), even though it is also possible to take advantage of common error cancellation in properly formulated undifferenced navigation schemes [32.92]. In either case, proper synchronization of measurements collected by receivers on individual spacecraft in a formation is a prerequisite for high-accuracy processing. Due to the high velocity of LEO satellites, even a subtle $1 \mu\text{s}$ epoch difference would show up as a 7 mm offset in the along-track component of the relative position. This raises the need for continuous clock steering within the employed GNSS receivers (Table 32.1) and poses specific constraints on the hardware selection for formation flying missions.

Considering the relevant contributions to the pseudorange and carrier-phase model introduced in Chap. 19, the single-difference $(\Delta(\cdot))_{ab} = (\cdot)_b - (\cdot)_a$ observations for two LEO spacecraft a and b tracking a common GNSS satellite s can be described as

$$\begin{aligned}\Delta p_{ab}^s &= \|\mathbf{r}^j - \mathbf{r}_b\| - \|\mathbf{r}^j - \mathbf{r}_a\| + c\Delta t_{ab} + \Delta I_{ab}^s, \\ \Delta \varphi_{ab}^s &= \|\mathbf{r}^j - \mathbf{r}_b\| - \|\mathbf{r}^j - \mathbf{r}_a\| + c\Delta t_{ab} - \Delta I_{ab}^s \\ &\quad + (\lambda \Delta N_{ab} + \Delta \delta_{ab}) + \lambda \Delta \omega_{ij}.\end{aligned}$$

(32.27)

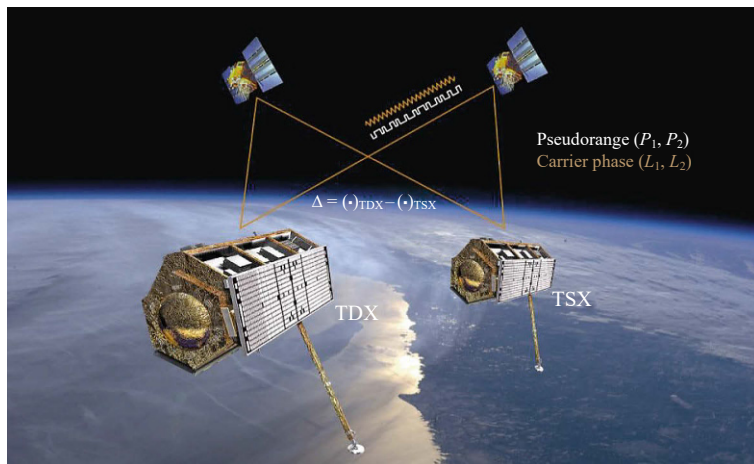


Fig. 32.15 Differential GPS observations for relative navigation in the TanDEM-X formation flying mission. Artist's drawing (courtesy of P. Kuss based on images of DLR/NASA)

Here \mathbf{r}^s and \mathbf{r}_i ($i = a, b$) denote the antenna position of the GNSS and LEO satellites at the signal transmission and reception time respectively, while Δt_{ab} is the single-difference of the receiver clock offsets. Differential ionospheric path delays ΔI_{ab}^s , if applicable, affect both code and phase observations but in an opposite manner, whereas differential phase ambiguities $\Delta \lambda N_{ab}$, receiver phase biases $\Delta \delta_{ab}$ and wind-up effects $\Delta \omega_{ab}$ are only present in the carrier-phase observation model.

Single differencing between two receivers rigorously eliminates GNSS clock offset uncertainties, which constitute a primary error source in single-spacecraft real-time navigation. The impact of GNSS position errors (and, equivalently, user position errors) on the computed relative position scales with the ratio of the baseline and the GNSS satellite distance and is thus attenuated by a factor of 100–10 000 in common formation flying missions. Even in onboard applications with representative broadcast ephemeris errors and absolute position uncertainties at the 1 m-level, the impact of such errors on the relative navigation remains below the carrier-phase noise for formations of up to 20 km separation.

The magnitude of differential ionospheric path delays for two orbiting receivers depends largely on the spacecraft separation but also on the total electron content above the spacecraft. While various efforts have been made to employ model-based ionosphere corrections in spaceborne relative navigation, only very limited practical knowledge of the differential ionosphere at orbital altitude and its variation with spacecraft separation is yet available from actual space missions. In the absence of more accurate models, differenced versions of the *Lear* model [32.61] presented in Sect. 32.2.2 have been employed to extend the range of single-frequency relative navigation [32.93] or to assist dual-frequency ambiguity resolution in long-baseline applications [32.94].

For the most common altitude range of 400–800 km, between 10 and 50% of the terrestrial vertical total electron content (VTEC) may be expected above a LEO satellite. During a phase of close proximity, differential ionospheric effects exceeding a few millimeters have been observed at baselines of more than about 10 km [32.95] for the GRACE satellites in late 2005. This supports the common conception that single-frequency relative navigation neglecting ionospheric errors is adequate for narrow formations with baselines of less than a few km. On the other hand, a rigorous elimination of ionospheric errors through dual-frequency linear combinations is clearly required for precise relative navigation at baselines of a few hundreds of kilometers as employed in the routine operations of the GRACE mission.

Similar to single-satellite precise orbit determination, dynamical models of the spacecraft motion in a relative navigation process are necessarily referred to the respective centers of gravity. Proper knowledge of the GNSS antenna phase-center relative to the center of gravity and concise information on the orientation of the spacecraft in space is therefore required for lever-arm correction in the observation modeling. However, GNSS antennas on board a spacecraft are commonly affected by phase-pattern distortions, which are caused by the near-field environment and hard to assess in pre-flight testing on ground. These distortions may reach a level of several millimeters even for geodetic-grade choke ring antennas (Fig. 32.10). An inflight calibration of absolute or relative antenna phase patterns is therefore essential to fully exploit the accuracy of differential GNSS carrier-phase observations in demanding formation flying applications [32.96].

Finally, differential phase wind-up effects $\Delta \omega_{ab}$ need to be considered in the most general case of a measurement model for differential carrier-phase observations in spacecraft relative navigation. Such effects show up whenever the boresight axes of the receiving antennas are not aligned with each other and do not coincide with the instantaneous rotation vector of the host spacecraft [32.97]. This situation is commonly encountered in rendezvous and proximity operations with specific constraints on the relative pose of the involved vehicles but mostly avoided in remote sensing formation flying missions with parallel alignment of all spacecraft.

While the use of receiver-receiver differences immediately simplifies the GNSS measurement model for a pair of formation flying spacecraft and helps to reduce or even fully eliminate common errors, it is less straightforward to apply the same concept to the relative dynamics. Considering a purely Keplerian motion (i. e., assuming a point-mass Earth), near-circular orbits, and close proximity, the relative motion of two spacecraft can be described by the Hill–Clohessy–Wiltshire equations [32.42] as a superposition of periodic oscillations in radial, along-track and cross-track direction as well as a linear drift in the along-track axis that is proportional to the mean radial separation. Since both spacecraft experience almost the same perturbations due to the asphericity of the Earth and the lunisolar gravity, it is tempting to neglect the resulting differential accelerations in the description of the relative motion. Even though such simplifications are adequate for conceptual studies of formation flying or orbit control purposes, they are not well suited to describe the relative dynamics at a level of accuracy compatible with GNSS carrier-phase observations. Also, drag and solar radiation pressure forces are unlikely to match exactly

for both spacecraft even in case of a similar design (e.g., due to different fuel loading or attitude). It is therefore advisable to employ a rigorous description of the orbital motion, which explicitly models all perturbations acting on the individual spacecraft.

Considering a dual-spacecraft formation, a combined state vector $\mathbf{y} = (\mathbf{r}_a^\top \mathbf{v}_a^\top \Delta \mathbf{r}_{ab}^\top \Delta \mathbf{v}_{ab}^\top)^\top$ can be formed from the absolute position and velocity of one spacecraft (here a) as well as the relative position and velocity of the two spacecraft. The corresponding equation of motion is given by

$$\frac{d\mathbf{y}}{dt} = \begin{pmatrix} \mathbf{v}_a \\ \mathbf{a}_a \\ \Delta \mathbf{v}_{ab} \\ \Delta \mathbf{a}_{ab} \end{pmatrix} \quad (32.28)$$

and requires modeling of the absolute acceleration for the reference satellite as well as the relative acceleration acting over the baseline of the formation. Depending on the separation of the two spacecraft, the relative acceleration model may employ some simplifications (such as a neglect of higher-order gravity field terms) but would nominally correspond to the difference of the absolute acceleration models for the two spacecraft. In addition, it may incorporate relative empirical accelerations that are typically required to compensate small deficiencies in the a priori model of the relative dynamics. On the other hand, (32.28) is computationally equivalent to the combination of the single-satellite states provided that the individual equations of motion are jointly integrated with a common stepsize. The latter approach is often preferred in practice, because it ensures a fully symmetric treatment of all spacecraft and can easily be generalized to multisatellite formations.

32.3.2 Estimation Concepts

The navigation problem for two or more spacecraft in a formation can be formulated in a variety of manners. The specific choice of estimation parameters may vary widely among different implementations and does not necessarily match the formulation of the dynamical state vector used in the equation of motion. Depending on the mission and application needs, a navigation process may be confined to the estimation of the relative state vector of a pair of satellites [32.98, 99], the absolute state of one spacecraft while keeping the state of the reference spacecraft fixed [32.100], or the absolute states of all individual members of the formation [32.101, 102]. In either case, knowledge of the relative motion is primarily derived from differential GNSS (carrier-phase) observations, while information on the absolute motion of the entire ensemble is, op-

tionally, provided from undifferenced single-satellite observations.

As discussed above, uncertainties in the absolute position of the reference spacecraft derived from a POD solution or a real-time navigation process are generally small enough to have negligible impact on the relative observation model. For all practical purposes it is therefore appropriate to keep the orbit of the reference satellite fixed at its a priori value and confine the estimation to the relative motion for each pair of satellites. In this way, the total number of estimation parameters can be minimized and the multisatellite adjustment can be partitioned into distinct estimation problems for the individual baselines.

In accord with the discussion of single-satellite navigation, extended Kalman filters are typically used for real-time onboard navigation systems, whereas the use of batch least-squares estimation techniques is confined to offline processing aiming at utmost accuracy. It may be noted, though, that an extended Kalman filter (EKF) filter/smoothing approach has also been preferred by [32.99] for the precise baseline reconstruction of dual-spacecraft formations such as GRACE or TanDEM-X. Benefits of the EKF design in this application include a notably reduced dimension of the estimation state and the possibility to resolve carrier-phase ambiguities on an epoch-by-epoch basis.

By way of example, the estimation state vector adopted in this particular relative navigation filter comprises the relative position $\Delta \mathbf{r}$ and velocity $\Delta \mathbf{v}$, relative drag and radiation pressure coefficients (ΔC_D , ΔC_R), relative empirical accelerations $\Delta \mathbf{a}_{\text{emp}}$ and the differential receiver clock offset $c\Delta t$, where spacecraft-related indices $(\cdot)_{ab}$ have been dropped for the ease of notation. For dual-frequency processing a differential ionospheric path delay ΔI^s as well as float-valued single-difference carrier-phase ambiguity parameters $\Delta A_j^s = \Delta N_j^s + \Delta \delta_j$ on both signal frequencies ($j = 1, 2$) are, furthermore, adjusted for each tracked satellite $s = (1, \dots, n)$. Overall, a 48-dimensional estimation vector is thus obtained for a 12-channel, single-constellation GNSS receiver. For single-frequency processing, the channel-wise ionospheric delays are replaced by a single vertical path delay parameter, which is used along with the Lear mapping function to compensate the differential path delays. Furthermore, only a single set of ambiguity parameters needs to be adjusted, which reduces the estimation state to 25 parameters.

The choice of estimation parameters illustrated here is representative for a single-baseline filter but in no way unique. Alternative parameterizations are discussed in the literature given above for different formulations of the relative navigation problem. In particular, the extension of a traditional POD concept for

precise baseline estimation using least-squares estimation is discussed in [32.100].

32.3.3 Ambiguity Resolution

Similar to other carrier-phase-based positioning applications, the resolution of integer ambiguities represents a key to utmost accuracy in spaceborne relative navigation. It effectively converts the ambiguous carrier-phase observations into low-noise pseudoranges and thus enables a mm-level relative positioning. While relative navigation solutions using float-ambiguity estimation are typically confined to accuracies in the 0.5–1 cm range, an up to ten times better performance has indeed been demonstrated for ambiguity-fixed solutions [32.99, 100].

A detailed presentation of carrier-phase ambiguity resolution and validation is given in Chap. 23. The present discussion is therefore confined to specific and practical aspects of ambiguity resolution in GNSS-based relative navigation of formation flying spacecraft. This application differs from terrestrial and airborne relative navigation in various aspects that are of relevance for ambiguity resolution:

- Continuous carrier-phase tracking arcs for a receiver in low Earth orbit last for typically less than 30 min and changes in the set of tracked satellites occur once every few minutes. Overall, some 500 single-difference ambiguities arise in a one-day dataset making it difficult to apply existing best integer estimation schemes for a joint resolution of all ambiguities.
- Due to limited communication bandwidths or restricted onboard processing capabilities that inhibit more frequent filter updates, the effective measurement rate is typically much less than in common real-time kinematic (RTK) applications. Considering, for example, a 30 s sampling interval, a spacecraft in low Earth orbit moves by roughly 200 km between two observations. This induces notable variations in the differential ionosphere for long baseline formations and affects both ambiguity resolution and cycle-slip detection capabilities in an adverse manner.
- As a positive aspect, the orbital motion causes a notable change of line-of-sight directions between the receiver and the tracked GNSS satellites. This results in an improved observability of both the relative position and the float ambiguity parameters. Further constraints are provided by the orbital dynamics, if the employed relative motion model is of adequate accuracy. Both aspects can assist the ambiguity resolution unless counteracted by lack-

ing knowledge of the differential ionospheric path delays.

In accordance with the properties of actual formation flying missions, ambiguity resolution has mainly been studied and applied for two extreme use cases. On the one side, narrow formations with baselines in the subkilometer range have been considered, where differential ionospheric path delays can essentially be neglected and relative navigation can readily be performed with single-frequency GNSS receivers. As shown in [32.98], ambiguity resolution can even be performed with a simplistic integer-rounding of float-valued double-difference ambiguities under these conditions. Large baselines of several hundreds of kilometers, in contrast, require use of dual-frequency observations and are substantially more challenging from an ambiguity resolution point of view. For batch least-squares estimation of the relative motion of the GRACE satellites, a bootstrapping approach has been adopted in [32.100]. Here double-difference wide-lane ambiguities are first determined based on the Melbourne–Wübbena combination of the dual-frequency observations, while the associated narrow-lane ambiguities are fixed from float ambiguity parameters estimated in the course of the relative orbit determination. As an alternative, the least-squares ambiguity decorrelation adjustment method (LAMBDA [32.103]) is employed in the relative navigation filter of [32.99], which implements a purely sequential processing scheme. Here, double differences are formed at each epoch from the single-difference float ambiguities estimated in the Kalman filter. A subset-fixing is then performed for reliably determined integer ambiguities. Both approaches have successfully been applied with actual flight data but depend on low-noise pseudorange observations to properly constrain the carrier-phase ambiguities in the absence of a suitable model for constraining the relative ionospheric path delays.

32.3.4 Flight Demonstrations

The use of GPS for relative navigation in space was first demonstrated in the mid-1990s as part of various rendezvous missions involving close approaches of the US space shuttle to free-flying payloads and the Russian Mir station [32.89, 104] and references therein). Most of these test campaigns were conducted in preparation for the European automated transfer vehicle (ATV), which makes use of GPS navigation during the far- and mid-range approach to the International Space Station (ISS) and conducted its maiden flight in 2008. Relative position accuracies of about 10 m and velocity accuracies of a few cm/s were achieved for ATV

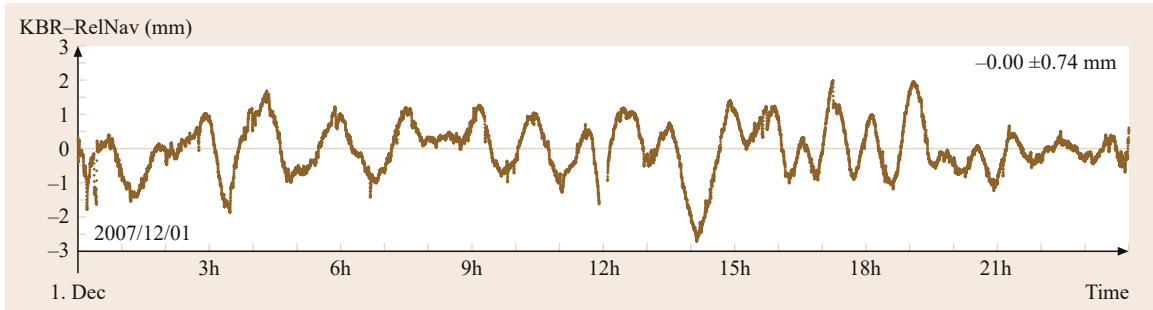


Fig. 32.16 Comparison of GPS-derived distances of the GRACE satellites with K-band ranging measurements (after [32.89])

and its precursor missions through filtering of differential pseudorange and range-rate observations along with simple models of the orbital dynamics. A similar performance has also been obtained in 1998 as part of the Japanese ETS-VII mission [32.105], which served as a platform for space robotics demonstration and conducted various GPS-controlled approaches of a chaser and target satellite.

However, none of the above mission were able to exploit the potential differential carrier-phase observations and the feasibility of mm-level relative navigation could only be verified many years later with flight data collected in the GRACE formation flying mission. GRACE comprises two identical spacecraft orbiting the Earth at a separation of about 200 km. Distance variations are continuously measured by a K-band intersatellite link with a precision of about $10\text{ }\mu\text{m}$. The formation thus acts as a large gradiometer and enables a detailed study of the Earth's gravity field and its temporal variations [32.106]. Due to the availability of geodetic-grade dual-frequency GPS receivers and the K-band ranging system, the GRACE formation represents a unique testbed for high-performance relative navigation and has triggered extensive research in the field [32.92, 99, 100]. Making use of high-fidelity a priori force models, empirical (relative) accelerations, integer ambiguity resolution, and in-flight phase pattern calibrations, a precision of 0.5–1.0 mm (1-D RMS) has consistently been demonstrated in comparison to K-band ranging measurements by various research groups (Fig. 32.16).

Building upon these pioneering results, relative GPS was later adopted for precision baseline reconstruction in the TanDEM-X SAR formation flying mission ([32.107], Fig. 32.15). The mission aims at the construction of a global digital elevation model (DEM) of unrivaled resolution and accuracy from interferometric SAR images. In order to avoid a tilt and shift of individual DEM tiles prior to the mosaicking, the relative position of the two spacecraft (or, more specifically,

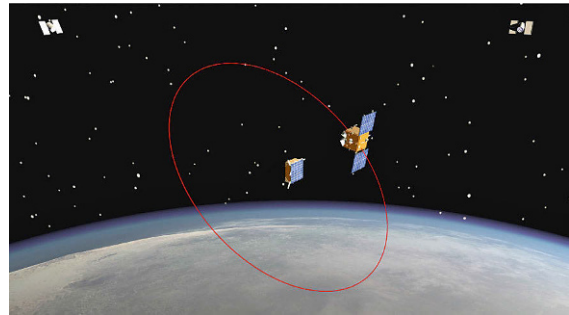


Fig. 32.17 GPS-enabled autonomous formation flying of the PRISMA satellites (courtesy of OHB Sweden)

the X-band SAR antennas) must be known with an accuracy of about 1 mm in both the radial and cross-track direction. Both satellites of the formation are therefore equipped with dual-frequency IGOR GPS receivers and choke ring antennas (Fig. 32.4) similar to those of the GRACE mission.

Other than GRACE, TanDEM-X offers no independent sensor system enabling a direct validation of the relative navigation accuracy. Various forms of consistency tests such as overlap tests, comparisons of single- and dual-frequency solutions, and, finally, comparisons of baseline products generated by different institutions using independent processing tools are therefore used to assess the resulting navigation performance. As demonstrated in [32.108], a consistency of 1 mm (1-D standard deviation) is typically achieved for baseline products generated by DLR, Deutsches GeoForschungsZentrum Potsdam (GFZ), and AIUB, but systematic biases of similar order may be noted. While slightly outside the mission specification this performance is still appropriate for DEM generation as demonstrated through dedicated tests with SAR calibration sites [32.109].

The use of carrier-phase differential GPS for accurate real-time navigation was first demonstrated in the Swedish PRISMA technology demonstration mis-

sion (Fig. 32.17), which was launched in 2010. The mission comprises two small satellites with a total mass of about 200 kg and serves as a testbed for autonomous formation flying and rendezvous operations. Both spacecraft are equipped with single-frequency Phoenix GPS receivers (Table 32.1). GPS pseudorange and carrier-phase observations collected on board the smaller target satellite are transmitted to the main spacecraft via a radio link and processed jointly with the local GPS observations in a real-time navigation system [32.102]. The GPS navigation system of PRISMA serves as the primary reference for monitoring the relative motion of the two spacecraft and enabled the first demonstration of autonomous formation control.

For maximum flexibility, the PRISMA navigation system processes both GRAPHIC observations of the individual satellites as well differential carrier-phase measurements of commonly observed GPS satellites. In

this way it can operate in all mission phases irrespective of the orientation of the two spacecraft and their GPS antennas. In view of frequent orbit and attitude maneuvers, priority is given to maximum robustness rather than utmost accuracy in the filter tuning. Also, the handling of carrier-phase ambiguities is limited to float ambiguity estimation and no effort is made to resolve integer ambiguities in the real-time onboard processing. Compared to a ground-based reference solution, the relative motion of the PRISMA could be determined on board with representative accuracies of better than 10 cm and 1 mm/s for position and velocity respectively [32.89, 102]. While this is still well below the theoretical performance limit for carrier-phase differential GPS, it is fully compatible with the PRISMA mission specification and represents a hundred times improvement over early relative navigation trials in the rendezvous missions discussed above.

32.4 Other Applications

The presentation given so far has focused on the use of GNSS for navigation and orbit determination of Earth orbiting satellites, which clearly represent the majority of applications for spaceborne GNSS receivers. The discussion would be incomplete, though, without addressing the use of GNSS for other types of space vehicles as well as nonnavigation-related applications. These topics are briefly introduced within this section.

32.4.1 Attitude Determination

The attractive performance of spaceborne GNSS receivers for positioning and navigation has also triggered an early interest in employing them for determining a satellite's attitude in space. Leaving aside single-antenna systems that exploit the directivity of the receiving antenna to obtain orientation-related information from measurements of the carrier-to-noise density ratio (see [32.110] and references therein), GNSS-based attitude determination relies on differential carrier-phase observations from typically three antennas forming two orthogonal baselines (Fig. 32.18). The respective principles and fundamental algorithms are presented in Chap. 27 in full detail. The present section is therefore confined to a short summary of relevant flight experiments and the experience gained in actual space missions.

Early demonstrations of GPS attitude determination were conducted on the RADCAL satellite [32.111, 112], the CRISTA-SPAS pallet satellite deployed by the US Space Shuttle [32.113] and the REX-II satel-

lite [32.114] in the 1993–1996 time frame. The experiments made use of Trimble TANS Vector or TANS Quadrex receivers that had been adapted for use in space and provided attitude information based on multiplexed carrier-phase measurements of four antennas. Since independent attitude sensors of adequate quality were not available on the above missions, the achieved performance has mainly been assessed through self-consistency tests or comparisons with low-grade references such as magnetometers. Typical accuracies of 0.5–1.0° have been reported, which is in rough accord

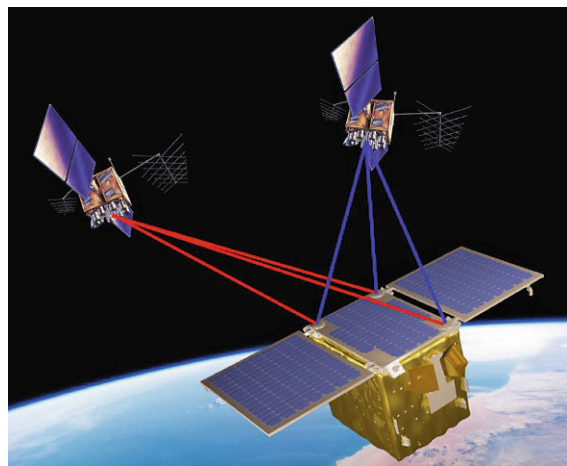


Fig. 32.18 GPS-based attitude determination of the *Flying Laptop* satellite (artist's drawing based on images of IFR, Univ. Stuttgart/NASA/US Gov)

with the performance expected for baselines of about 0.5 m and carrier-phase accuracies at the level of few millimeters.

Within Europe several demonstrations of GPS-based attitude determination have been conducted on various SSTL microsatellites equipped with SGR-20 four-antenna GPS receivers. Typical accuracies of about 1–2° have, for example, been obtained for UoSAT-12 and TopSat [32.115, 116] in comparison with horizon sensors providing a reference attitude at the 0.2°-level.

Finally, GPS measurements from the Space Integrated GPS/Inertial navigation system (SIGI) are routinely used for attitude determination system on board the International Space Station (ISS). The SIGI unit comprises a Trimble Force 19 GPS receiver connected to an array of four antennas on the S0 truss of the ISS [32.117]. Despite the use of multipath limiting choke ring antennas and an antenna baseline of several meters, the achieved standalone GPS attitude performance does not meet the 0.5° specification. GPS-based attitude solutions are therefore merged with other sensor data in the onboard attitude determination filter to obtain a robust and accurate blended solution.

Overall, multipath and phase pattern distortions remain a limiting factor for the achievable accuracy for GNSS-based attitude determination on space vehicles. Surface space is naturally limited on common satellites, which inhibits the accommodation of large and widely separated antennas. Despite initial hopes, GNSS attitude sensors have not been found to be competitive with Earth horizon sensors or star cameras and their use has mainly been confined to selected experiments and demonstration campaigns.

32.4.2 Ballistic Missions

Aside from Earth orbiting satellites, GNSS is also an attractive tracking system for ballistic space vehicles that stay in space for only a limited time before returning to ground. Among others, GNSS receivers are now widely used on sounding rockets, which provide low-cost platforms for atmospheric and astronomical research as well as biological and physical experiments in microgravity. Sounding rockets fly at suborbital velocities ($v < 7$ km) but may reach altitudes of 100–1000 km during their parabolic flight phase.

As an example, Fig. 32.19 illustrates the flight parameters of a Maxus rocket, which represents the most powerful sounding rocket employed in the European microgravity research program. Within the 60 s boost phase the vehicle achieves a total speed of 3200 m/s and is accelerated with about 100 m/s (10 G) near burnout of the second stage. Even higher accelera-

tions of 50 G and more are experienced near the end of the 900 s flight when the payload enters the atmosphere.

GNSS tracking of sounding rockets is primarily employed for flight safety purposes and complements or substitutes ground-based radars for the instantaneous impact point (IIP) prediction during the propelled flight. The IIP describes the touchdown point reached by the vehicle as computed from the current position and velocity [32.118]. It is used to validate a nominal flight performance during the boost phase and, if necessary, to destroy the booster in case of anomalies. Other uses of GNSS navigation data may include onboard timing, relative positioning of ejected payloads [32.119], payload recovery, and finally overall flight performance assessment.

The extreme dynamics of a sounding rocket flight poses specific requirements on the environmental robustness and the signal processing within a GNSS receiver. The employed hardware must be robust enough to withstand high levels of vibration and to operate in vacuum (but need to be radiation-hardened due to the very short mission duration). To cope with high accelerations and acceleration changes (jerk), the code- and carrier-tracking loops (Chap. 14) must be of adequate order and width. In addition high navigation update rates (5–20 Hz) are typically desired to resolve individual flight phases with adequate resolution. Antenna accommodation represents a further challenge for the use of GNSS on sounding rockets. Wraparound antennas are commonly favored to enable continuous tracking irrespective of vehicle spin and attitude but need to be tailor-made for the specific body diameter and are unsuitable for large vehicles. Alternative concepts offering different levels of complexity and reliability include helical tip antennas as well as body-mounted blade and tip antennas.

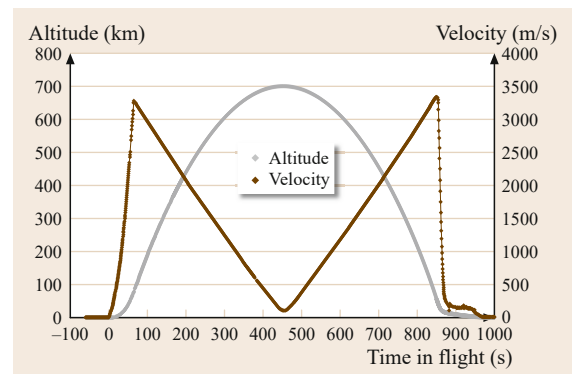


Fig. 32.19 Altitude and total velocity as measured by an Orion GPS receiver during the Maxus-5 sounding rocket flight in April 2003

In terms of performance, the better than 10 m and 0.1 m/s navigation accuracy offered by public GNSS signals and services is generally deemed fully adequate for sounding rockets and other (nonmilitary) ballistic vehicles. In accord with this, single-frequency GNSS systems are almost exclusively employed in view of the reduced antenna and receiver complexity. While external reference standards are rarely available, a comparison of multiple GPS L1 C/A-code receivers and antenna assemblies flown jointly on an Orion sounding rocket in 2001 [32.120] has indeed confirmed the desired precision for most of the employed systems.

Despite an adequate navigation performance GNSS cannot normally ensure a fully continuous availability as a standalone sensor but is subject to a potential loss-of-track in phases of extreme acceleration changes, antenna switches, unfavorable attitude, rapid spin or blackouts due to ionization. Also, strong scintillation is considered a potential risk for use at launch sites in tropical regions. A coupling of GNSS sensors with inertial navigation systems (INSs) is therefore considered mandatory for use on guided vehicles, large launchers, human space flight and re-entry capsules [32.121, 122].

An example of an advanced GPS/INS navigation system for sounding rocket missions is shown in Fig. 32.20. The hybrid navigation system [32.123] was first flown as part of the Shefex-II sounding rocket mission and employs a tightly coupled system architecture (Chap. 28). It computes a strapdown navigation solution from high-rate (400 Hz) acceleration and angular rate observations of an Inertial Measurement Unit (IMU) with three servo accelerometers and fiber optic gyros. This is updated once per second in a common navigation filter with raw pseudorange observations of a GPS L1 C/A-code receiver.

32.4.3 GNSS Radio Science

Beyond their primary purpose, GNSS navigation signals may also serve as signals of opportunity for different types of remote sensing applications from spaceborne platforms. This includes their use in radio occultation measurements for sounding the state of

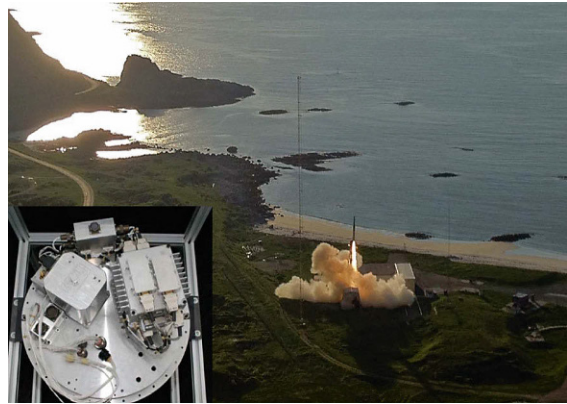


Fig. 32.20 Launch of the SHEFEX-2 rocket from Andoya rocket base and hybrid navigation system (courtesy of DLR)

the troposphere and ionosphere [32.75] but also scatterometry and reflectometry for surveying the ocean surface [32.124]. While the first application is already supported by numerous space missions (such as CHAMP, Metop, TerraSAR-X and COSMIC) and has in fact evolved into an indispensable data source for near-real-time weather forecasts [32.125], the use of reflected or scattered GNSS signals from space is still in an experimental state [32.126] and waiting for dedicated missions and payloads to be developed.

GNSS receivers for radio occultation (RO) observations are offered by various manufacturers building upon existing concepts of spaceborne dual-frequency receivers for precise orbit determination. Typical differences include the need for multiple antennas (forward/backward looking for RO, zenith looking for POD), open-loop tracking with model-based predictions of the expected code delay and high-data rate phase observations (or raw in-phase/quadrature (I/Q) samples). Furthermore, dedicated antenna arrays with beam patterns focused on the limb of the Earth are commonly used to compensate the attenuation of signals passing deep through the lower atmosphere.

For a more detailed discussion of GNSS radio occultation as well reflectometry and scatterometry the reader is referred to Chapters 38 and 40 respectively.

References

- 32.1 J. Rush: Current issues in the use of the Global Positioning System aboard satellites, *Acta Astronaut.* **47**(2–9), 377–387 (2000)
- 32.2 O. Montenbruck, M. Markgraf, M. Garcia, A. Helm: GPS for microsatellites – status and perspectives. In: *Small Satellites for Earth Observation*, ed. by R. Sandau, H.P. Röser, A. Valenzuela (Springer, Heidelberg 2008) pp. 165–174
- 32.3 J.R. Vetter: Fifty years of orbit determination: Development of modern astrodynamics methods, *John Hopkins APL Tech. Dig.* **27**(3), 239–252 (2007)

- 32.4 W.P. Birmingham, B.L. Miller, W.L. Stein: Experimental results of using the GPS for Landsat-4 onboard navigation, *Navigation* **30**(3), 244–251 (1983)
- 32.5 S.C. Wu, T.P. Yunck, C.L. Thornton: Reduced-dynamic technique for precise orbit determination of low Earth satellites, *J. Guid. Control Dyn.* **14**(1), 24–30 (1991)
- 32.6 T.P. Yunck, W.L. Bertiger, S.C. Wu, Y.E. Bar-Sever, E.J. Christensen, B.J. Haines, S.M. Lichten, R.J. Muellerschoen, E.S. Davis, J.R. Guinn, Y. Vigue, P. Willis: First assessment of GPS-based reduced dynamic orbit determination on TOPEX/Poseidon, *Geophys. Res. Letters* **21**(7), 541–544 (1994)
- 32.7 R. Ware, C. Rocken, F. Solheim, M. Exner, W. Schreiner, R. Anthes, D. Feng, B. Herman, M. Gorbunov, S. Sokolovskiy, K. Hardy, Y. Kuo, X. Zou, K. Trenberth, T. Meehan, W. Melbourne, S. Businger: GPS sounding of the atmosphere from low Earth orbit: Preliminary Results, *Bull. Am. Meteorological Soc.* **77**, 19–40 (1996)
- 32.8 W. Marquis: The GPS Block IIR/IIR-M antenna panel pattern. Lockheed Martin Corp. (2014) <http://www.lockheedmartin.com/us/products/gps/gps-publications.html>
- 32.9 Th.D. Powell, Ph.D. Martzen, S.B. Sedlacek, C.-C. Chao, R. Silva, A. Brown, G. Belle: GPS signals in a geosynchronous transfer orbit: Falcon Gold data processing, *Proc. ION ITM* (1999) pp. 575–585
- 32.10 O. Balbach, B. Eissfeller: Analyses of the Equator-S GPS Mission Data at Altitudes above the GPS-Constellation, *Proc. 4th ESA Intern. Conf. on Spacecr. Guidance, Navigation and Control Systems*, Noordwijk, ed. by B. Schürmann (ESA, Netherlands 2000) pp. 131–137
- 32.11 M.C. Moreau, E.P. Davis, J.R. Carpenter, D. Kelbel, G.W. Davis, P. Axelrad: Results from the GPS flight experiment on the high Earth orbit AMSAT OSCAR-40 spacecraft, *Proc. ION GPS* (2002) pp. 122–133
- 32.12 J.D. Kronman: Experience using GPS for orbit determination of a geosynchronous satellite, *Proc. ION GPS* (2000) pp. 1622–1626
- 32.13 M. Unwin, R. De Vos Van Steenwijk, P. Blunt, Y. Hashida, S. Kowaltschek, L. Nowak: Navigating above the GPS constellation – Preliminary results from the SGR-GEO on GIOVE-A, *Proc. ION GNSS* (2013) pp. 3305–3315
- 32.14 L.M.B. Winternitz, W.A. Bamford, G.W. Heckler: A GPS receiver for high-altitude satellite navigation, *IEEE J. Sel. Top. Signal Process.* **3**(4), 541–556 (2009)
- 32.15 M. Farahmand, A. Long, R. Carpenter: Magnetospheric multiscale mission navigation performance using the goddard enhanced onboard navigation system, *Proc. Int. Symp. Space Flight Dynamics ISSFD*, München ed. by R. Kahle (DLR, Oberpfaffenhofen 2015) pp. 1–17
- 32.16 F.H. Bauer, M.C. Moreau, M.E. Dahle-Melsaether, W.P. Petrofski, B.J. Stanton, S. Thomason, G.A. Harris, R.P. Sena, L.P. Temple III: The GPS space service volume, *Proc. ION GNSS* (2006) pp. 2503–2514
- 32.17 J.P.W. Stark: The spacecraft environment and its effect on design. In: *Spacecraft Systems Engineering*, ed. by P. Fortescue, G. Swinerd, J. Stark (Wiley, New York 2011) pp. 11–48
- 32.18 J.M. Rabaey, A. Chandrakasan, B. Nikolic: *Digital Integrated Circuits*, 2nd edn. (Prentice Hall, New Jersey 2002)
- 32.19 J. Roselló, P. Silvestrin, R. Weigand, S. d’Addio, A. García-Rodríguez, G. López Risueño: Next Generation of ESA’s GNSS Receivers for Earth Observation Satellites, *NAVITEC’2012*, Noordwijk, Netherlands (IEEE, 2012) pp. 1–8
- 32.20 M.J. Unwin, M.K. Oldfield: The Design and Operation of a COTS Space GPS Receiver, *Proc. 23rd Annual AAS Guidance & Control Conference*, Breckenridge (2000) pp. 00–046
- 32.21 A. Hauschild, M. Markgraf, O. Montenbruck: Flight results of the NOX dual-frequency GPS receiver payload on-board the TET satellite, *Proc. ION GNSS* (2013) pp. 316–3324
- 32.22 A. Helm, M.-P. Hess, M. Minori, A. Gribkov, S. Yudanov, O. Montenbruck, G. Beyerle, L. Cacciapuoti, R. Nasca: The ACES GNSS subsystem and its potential for radio-occultation and reflectometry from the international space station, *Proc. 2nd Int. Colloquium on Scientific and Fundam. Aspects of the Galileo Program*, Padua (2009)
- 32.23 E. Kahr, O. Montenbruck, K. O’Keefe, S. Skone, J. Urbanek, L. Bradbury, P. Fenton: GPS tracking of a nanosatellite – The CanX-2 flight experience, *Proc. 8th Int. ESA Conf. Guidance, Navigation & Control Systems*, Carlsbad (ESA, Noordwijk 2010) pp. 1–13
- 32.24 European Commission: Council Regulation (EC) No 428/2009 of 5 May 2009 setting up a Community regime for the control of exports, transfer, brokering and transit of dual-use items (European Commission, Brussels 2009) http://trade.ec.europa.eu/doclib/docs/2009/june/tradoc_143390.pdf
- 32.25 US Department of State: *International Traffic in Arms Regulations 2011* (US Department of State, Directorate of Defense Trade Controls 2011) http://www.pmdtc.state.gov/regulations_laws/itar.html
- 32.26 H. Bock, U. Hugentobler, G. Beutler: Kinematic and dynamic determination of trajectories for low Earth satellites using GPS. In: *First CHAMP Mission Results for Gravity, Magnetic and Atmospheric Studies*, ed. by Ch. Reigber, H. Lühr, P. Schwintzer (Springer, Berlin 2003) pp. 65–69
- 32.27 D. Švehla, M. Rothacher: Kinematic precise orbit determination for gravity field determination. In: *A Window on the Future of Geodesy*, ed. by F. Sansò (Springer, Berlin 2005) pp. 181–188
- 32.28 O. Montenbruck, E. Gill: *Satellite Orbits – Models, Methods and Applications* (Springer, Berlin 2005)
- 32.29 E. Hairer, S.P. Norsett, G. Wanner: *Solving Ordinary Differential Equations I. Nonstiff Problems* (Springer, Berlin 1987)
- 32.30 G. Beutler: *Methods of Celestial Mechanics* (Springer, Berlin 2005)

- 32.31 E. Fehlberg: Classical Fifth-, Sixth-, Seventh-, and Eight-Order Runge-Kutta Formulas with Stepsize Control (NASA, Washington DC 1968)
- 32.32 P.J. Prince, J.R. Dormand: High order embedded Runge-Kutta formulae, *J. Comp. Appl. Math.* **7**, 67–75 (1981)
- 32.33 L.F. Shampine, M.K. Gordon: *Computer Solution of Ordinary Differential Equations* (Freeman and Comp., San Francisco 1975)
- 32.34 M.M. Berry: A Variable-Step Double-Integration Multi-Step Integrator, Ph.D. Thesis (Virginia Polytechnic Institute and State University, Blacksburg 2004)
- 32.35 D.E. Pavlis, S.G. Poulos, C. Deng, J.J. McCarthy: GEODYN II System Documentation, SGT-Inc., Greenbelt, MD, contractor report (2007)
- 32.36 T. Springer: *NAPEOS Mathematical Models and Algorithms* (ESA, Darmstadt 2009)
- 32.37 S.M. Lichten, Y.E. Bar-Sever, W.I. Bertiger, M. Heflin, K. Hurst, R.J. Muellerschoen, S.C. Wu, T.P. Yunck, J. Zumbeke: Gipsy-Oasis II: A high precision GPS data processing system and general satellite orbit analysis tool, *Technology*, 24–26 (2005)
- 32.38 M. Wermuth, O. Montenbruck, T. van Helleputte: GPS high precision orbit determination software tools (GHOST), Proc. 4th Int. Conf. Astrodyn. Tools Tech., Madrid (ESA, Noordwijk 2010)
- 32.39 R. Dach, U. Hugentobler, P. Fridez, M. Meindl: *Bernese GPS Software, Software Version 5.0* (Astronomical Institute University of Bern, Switzerland 2007)
- 32.40 O. Montenbruck, E. Gill: State interpolation for on-board navigation systems, *Aerosp. Sci. Technol.* **5**(3), 209–220 (2001)
- 32.41 E. Gill, O. Montenbruck, H. Kayal: The BIRD satellite mission as a milestone toward GPS-based autonomous navigation, *Navigation* **48**(2), 69–76 (2001)
- 32.42 D.A. Vallado: *Fundamentals of Astrodynamics and Applications*, 2nd edn. (Kluwer Academic, Dordrecht 2001)
- 32.43 A. Milani, A.M. Nobili, P. Farinella: *Non-Gravitational Perturbations and Satellite Geodesy* (Adam Hilger, Bristol 1987)
- 32.44 L.E. Cunningham: On the computation of the spherical harmonic terms needed during the numerical integration of the orbital motion of an artificial satellite, *Celest. Mech.* **2**(2), 207–216 (1970)
- 32.45 D. Tsoulis, K. Patlakis: A spectral assessment review of current satellite-only and combined earth gravity models, *Rev. Geophys.* **51**(2), 186–243 (2013)
- 32.46 J. Bouman, R. Floberghagen, R. Rummel: More than 50 years of progress in satellite gravimetry, *EOS Trans. Am. Geophys. Union* **94**(31), 269–270 (2013)
- 32.47 Ch.M. Botai, L. Combrinck: Global geopotential models from satellite laser ranging data with geophysical applications: A review, *South African J. Sci.* **108**(3/4), 1–10 (2012)
- 32.48 H. Klinkrad, B. Fritsche: Orbit and attitude perturbations due to aerodynamics and radiation pressure, Proc. ESA Workshop on Space Weather, Noordwijk (ESA, Noordwijk 1998), NL 1998
- 32.49 E. Doornbos: Thermospheric Density and Wind Determination from Satellite Dynamics, Ph.D. Thesis (Tu Delft, Delft 2012)
- 32.50 J.M. Picone, A.E. Hedin, D.P. Drob, A.C. Aikin: NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues, *J. Geophys. Res.* **107**(A12), SIA 15/1–SIA 15/16 (2002)
- 32.51 S.L. Bruinsma, N. Sánchez-Ortiz, E. Olmedo, N. Guijarro: Evaluation of the DTM-2009 thermosphere model for benchmarking purposes, *J. Space Weather Space Clim.* **2**, A04 (2012)
- 32.52 D.A. Vallado, D. Finkleman: A critical assessment of satellite drag and atmospheric density modeling, Proc. AIAA 2008–6442, AIAA /AAS Astrodynamics Specialist Conference, Honolulu (AIAA, Reston 2008) pp. 1–28
- 32.53 A.J. Sibthorpe: Precision Non-conservative Force Modelling for Low Earth Orbiting Spacecraft, Ph.D. Thesis (University of London, London 2006)
- 32.54 G. Petit, B. Luzum: *IERS Conventions 2010*, IERS Technical Note No. 36 (Bundesamt für Kartographie und Geodäsie, Frankfurt am Main 2010)
- 32.55 M. Eineder, N. Adam, R. Bamler, N. Yague-Martinez, H. Breit: Spaceborne spotlight SAR interferometry with TerraSAR-X, geoscience and remote sensing, *IEEE Trans.* **47**(5), 1524–1535 (2009)
- 32.56 M. Grondin, J.L. Issler, M.C. Charneau, D. Lamy, A. Laurichesse, P. Raizonville, M.-A. Clair, C. Mehlen, C. Boyer, N. Wilhelm, H. Favaro: Autonomous orbit control with GPS on board the DEMETER spacecraft, Proc. NAVITEC, Noordwijk (ESA, Noordwijk 2006)
- 32.57 P.C.P.M. Pardal, H.K. Kuga, R.V. de-Morales: Comparing the extended and the sigma point Kalman filters for orbit determination modeling using GPS measurements, Proc. ION GNSS (2010) pp. 2732–2742
- 32.58 B. Tapley, B. Schutz, G.H. Born: *Statistical Orbit Determination* (Elsevier, Amsterdam 2004)
- 32.59 G.J. Bierman: *Factorization Methods for Discrete Sequential Estimation* (Courier Dover, Mineola 2006)
- 32.60 J.A. Klobuchar: Ionospheric time-delay algorithm for single-frequency GPS users, *IEEE Trans. Aerosp. Electron. Syst.* **23**(3), 325–331 (1987)
- 32.61 W.M. Lear: *GPS Navigation for Low-Earth Orbiting Vehicles*, NASA 87-FM-2, Rev. 1, JSC-32031 (Lyndon B. Johnson Space Center, Houston 1987)
- 32.62 O. Montenbruck, P. Ramos-Bosch: Precision real-time navigation of LEO satellites using global positioning system measurements, *GPS Solut.* **12**(3), 187–198 (2008)
- 32.63 B. Gruber: GPS program update, Proc. ION GNSS (2012) pp. 521–537
- 32.64 C. Jayles, J.P. Chauveau, F. Roza: DORIS/Jason-2: Better than 10cm on-board orbits available for near-real-time altimetry, *Adv. Space Res.* **46**(12), 1497–1512 (2010)

- 32.65 A. Reichert, T. Meehan, T. Munson: Toward decimeter-level real-time orbit determination: A demonstration using the SAC-C and CHAMP spacecraft, *Proc. ION GPS* (2002) pp. 1996–2003
- 32.66 Y. Bar-Sever, L. Young, F. Stocklin, P. Heffernan, J. Rush: The NASA global differential GPS system (GDGPS) and the TDRSS augmentation service for satellites (TASS), *Proc. NAVITEC*, Noordwijk (ESA, Noordwijk 2004) pp. 1–8
- 32.67 M. Saito, Y. Sato, M. Miya, M. Shima, Y. Omura, J. Takiguchi, K. Asari: Centimeter-class augmentation system utilizing quasi-zenith satellite, *Proc. ION GNSS* (2011) pp. 1243–1253
- 32.68 T.P. Yunck: Orbit determination. In: *Global Positioning System – Theory and Applications*, ed. by B.W. Parkinson, J.J. Spilker (AIAA, Washington DC 1996)
- 32.69 O. Montenbruck, P. Swatschina, M. Markgraf, S. Santandrea, J. Naudet, E. Tilmans: Precision spacecraft navigation using a low-cost GPS receiver, *GPS Solut.* **16**(4), 519–529 (2012)
- 32.70 B.D. Beckley, N.P. Zelensky, S.A. Holmes, F.G. Lemoine, R.D. Ray, G.T. Mitchum, S.D. Desai, S.T. Brown: Assessment of the Jason-2 extension to the TOPEX/Poseidon, Jason-1 sea-surface height time series for global mean sea level monitoring, *Mar. Geod.* **33**(S1), 447–471 (2010)
- 32.71 PODAAC: Global Mean Sea Level Trend from Integrated Multi-Mission Ocean Altimeters TOPEX/Poseidon Jason-1 and OSTM/Jason-2 Version 2. PO.DAAC, CA, USA. Dataset accessed 26 Oct. 2015 at <http://dx.doi.org/10.5067/GMSLM-TJ122>
- 32.72 H. Breit, Th. Fritz, U. Balss, M. Lachaise, A. Niedermeier, M. Vonavka: TerraSAR-X SAR processing and products, *IEEE Trans. Geosci. Remote Sens.* **48**(2), 727–740 (2010)
- 32.73 M. Eineder, Ch. Minet, P. Steigenberger, X. Cong, Th. Fritz: Imaging geodesy – Toward centimeter-level ranging accuracy with TerraSAR-X, geoscience and remote sensing, *IEEE Trans.* **49**(2), 661–671 (2011)
- 32.74 L. Cerri, J.P. Berthias, W.I. Bertiger, B.J. Haines, F.G. Lemoine, F. Mercier, J.C. Ries, P. Willis, N.P. Zelensky, M. Ziebart: Precision orbit determination standards for the Jason series of altimeter missions, *Mar. Geod.* **33**(S1), 379–418 (2010)
- 32.75 E.R. Kursinski, G.A. Hajj, J.T. Schofield, R.P. Linfoot, K.R. Hardy: Observing Earth's atmosphere with radio occultation measurements using the Global Positioning System, *J. Geophys. Res.* **102**(D19), 23429–23465 (1997)
- 32.76 O. Montenbruck, M. Wermuth, A. Hauschild, G. Beyerle, A. Helm, S. Yudanov, A. Garcia, L. Cacciapuoti: Multi-GNSS precise orbit determination of the International Space Station, *Proc. ION ITM* (2013) pp. 808–820
- 32.77 O. Montenbruck, T. Van-Helleputte, R. Kroes, E. Gill: Reduced dynamic orbit determination using GPS code and carrier measurements, *Aerosp. Sci. Technol.* **9**(3), 261–271 (2005)
- 32.78 O. Montenbruck, Y. Andres, H. Bock, T. van-Helleputte, J. van-den-Ijssel, M. Loiselet, C. Marquardt, P. Silvestrin, P. Visser, Y. Yoon: Tracking and orbit determination performance of the GRAS instrument on Metop-A, *GPS Solut.* **12**(4), 289–299 (2008)
- 32.79 M.R. Pearlman, J.J. Degnan, J.M. Bosworth: The International Laser Ranging Service, *Adv. Space Res.* **30**(2), 135–143 (2002)
- 32.80 H. Bock, A. Jäggi, U. Meyer, P. Visser, J. van-den-Ijssel, T. van Helleputte, M. Heinze, U. Hugentobler: GPS-derived orbits for the GOCE satellite, *J. Geodesy* **85**(11), 807–818 (2011)
- 32.81 C. Flohrer, M. Otten, T. Springer, J. Dow: Generating precise and homogeneous orbits for Jason-1 and Jason-2, *Adv. Space Res.* **48**(1), 152–172 (2011)
- 32.82 W. Bertiger, S.D. Desai, A. Dorsey, B.J. Haines, N. Harvey, D. Kuang, A. Sibthorpe, J.P. Weiss: Sub-centimeter precision orbit determination with GPS for ocean altimetry, *Mar. Geod.* **33**(S1), 363–378 (2010)
- 32.83 D. Laurichesse, F. Mercier, J.P. Berthias, P. Broca, L. Cerri: Integer ambiguity resolution on undifferenced GPS phase measurements and its application to PPP and satellite precise orbit determination, *Navigation* **56**(2), 135 (2009)
- 32.84 W. Bertiger, S.D. Desai, B. Haines, N. Harvey, A.W. Moore, S. Owen, J.P. Weiss: Single receiver phase ambiguity resolution with GPS data, *J. Geodesy* **84**(5), 327–337 (2010)
- 32.85 R. Zandbergen, A. Ballereau, E. Rojo, Y. Andres, I. Romero, C. Garcia, J.M. Dow: GRAS GSN near-real time data processing, *Proc. IGS Workshop*, Darmstadt (IGS, Pasadena 2006) pp. 1–20, 8–12 May 2006
- 32.86 M. Caissy, L. Agrotis, G. Weber, M. Hernandez-Pajares, U. Hugentobler: Coming soon – The international GNSS real-time service, *GPS World* **23**(6), 52 (2012)
- 32.87 O. Montenbruck, A. Hauschild, Y. Andres, A. von Engel, Ch. Marquardt: (Near-) real-time orbit determination for GNSS radio occultation processing, *GPS Solut.* **17**(2), 199–209 (2013)
- 32.88 B.J. Haines, M.J. Armatus, Y.E. Bar-Sever, W.I. Bertiger, S.D. Desai, A.R. Dorsey, Ch.M. Lane, J.P. Weiss: One-centimeter orbits in near-real time: The GPS experience on OSTM/JASON-2, *J. Astronaut. Sci.* **58**(3), 445–459 (2011)
- 32.89 O. Montenbruck, S. D'Amico: GPS based relative navigation. In: *Distributed Space Missions for Earth System Monitoring*, ed. by M. D'Errico (Springer, New York 2013) pp. 185–223
- 32.90 M. D'Errico: *Distributed Space Missions for Earth System Monitoring* (Springer, Berlin 2013)
- 32.91 K. Alfriend, S.R. Vadali, P. Gurfil, J. How, L. Breger: *Spacecraft Formation Flying: Dynamics, Control, and Navigation* (Butterworth-Heinemann, Oxford 2010)
- 32.92 S.C. Wu, Y.E. Bar-Sever: Real-time sub-cm differential orbit determination of two low-Earth orbiters with GPS bias Fixing, *Proc. ION GNSS* (2006) pp. 2515–2522
- 32.93 O. Montenbruck, M. Wermuth, R. Kahle: GPS based relative navigation for the TanDEM-X Mis-

- sion – First flight results, *Navigation* **58**(4), 293–304 (2011)
- 32.94 U. Tancredi, A. Renga, M. Grassi: Ionospheric path delay models for spaceborne GPS receivers flying in formation with large baselines, *Adv. Space Res.* **48**(3), 507–520 (2011)
- 32.95 P.W.L. van Barneveld, O. Montenbruck, P.N.A.M. Visser: Epochwise prediction of GPS single differenced ionospheric delays of formation flying spacecraft, *Adv. Space Res.* **44**(9), 987–1001 (2009)
- 32.96 A. Jäggi, R. Dach, O. Montenbruck, U. Hugentobler, H. Bock, G. Beutler: Phase center modeling for LEO GPS receiver antennas and its impact on precise orbit determination, *J. Geodesy* **83**(12), 1145–1162 (2009)
- 32.97 M.L. Psiaki, S. Mohiuddin: Modeling, analysis, and simulation of GPS carrier phase for spacecraft relative navigation, *J. Guid. Control Dyn.* **30**(6), 1628–1639 (2007)
- 32.98 S. Leung, O. Montenbruck: Real-time navigation of formation-flying spacecraft using global-positioning-system measurements, *J. Guid. Control Dyn.* **28**(2), 226–235 (2005)
- 32.99 R. Kroes: Precise Relative Positioning of Formation Flying Spacecraft using GPS, Ph.D. Thesis (TU Delft, Delft 2006)
- 32.100 A. Jäggi, U. Hugentobler, H. Bock, G. Beutler: Precise orbit determination for GRACE using undifferenced or doubly differenced GPS data, *Adv. Space Res.* **39**(10), 1612–1619 (2007)
- 32.101 T. Ebinuma, R.H. Bishop, E.G. Lightsey: Spacecraft rendezvous using GPS relative navigation, *Proc. AAS 01-152, AAS/AIAA Space Flight Mechanics Meeting*, Santa Barbara (2001) pp. 701–718
- 32.102 S. D’Amico, J.S. Ardaens, R. Larsson: Spaceborne autonomous formation-flying experiment on the PRISMA mission, *J. Guid. Control Dyn.* **35**(3), 834–850 (2012)
- 32.103 P.J.G. Teunissen: The least-squares ambiguity decorrelation adjustment: A method for fast GPS integer ambiguity estimation, *J. Geodesy* **70**(1), 65–82 (1995)
- 32.104 G. Moreau, H. Marcollé: RGPS post-flight analysis of ARP-K flight demonstration, *Proc. Int. Symp. Space Flight Dynamics*, Darmstadt (ESA SP403, Noordwijk 1997) pp. 97–102
- 32.105 I. Kawano, M. Mokuo, T. Kasai, T. Suzuki: First autonomous rendezvous using relative GPS navigation by ETS-VII, *Navigation* **48**(1), 49–56 (2001)
- 32.106 B.D. Tapley, S. Bettadpur, M. Watkins, Ch. Reigber: The gravity recovery and climate experiment: Mission overview and early results, *Geophys. Res. Letters* **31**(L0960), 1–4 (2004)
- 32.107 G. Krieger, A. Moreira, H. Fiedler, I. Hajnsek, M. Werner, M. Younis, M. Zink: TanDEM-X: A satellite formation for high-resolution SAR interferometry, *IEEE Trans. Geosci. Remote Sens.* **45**(11), 3317–3341 (2007)
- 32.108 A. Jäggi, O. Montenbruck, Y. Moon, M. Wermuth, R. König, G. Michalak, H. Bock, D. Bodenmann: Inter-agency comparison of TanDEM-X baseline solutions, *Adv. Space Res.* **50**(2), 260–271 (2012)
- 32.109 M. Wermuth, O. Montenbruck, A. Wendleder: Relative navigation for the TanDEM-X mission and evaluation with DEM calibration results, *J. Aerosp. Eng.* **3**(2), 28–38 (2011)
- 32.110 C. Wang, R. Walker, M. Moody: An improved single antenna attitude system based on GPS signal strength, *Proc. AIAA 2005-5993, AIAA Guidance, Navigation, and Control Conference and Exhibit*, San Francisco (AIAA, Reston 2005) pp. 1–15
- 32.111 E.G. Lightsey, C.E. Cohen, B.W. Parkinson: Attitude determination and control for spacecraft using differential GPS, *Proc. Int. Conf. on GNC, Noordwijk (ESA WPP-071, Noordwijk 1994)* pp. 453–461
- 32.112 P. Axelrad, L.M. Ward: Spacecraft attitude estimation using the global positioning system – Methodology and results for RADCAL, *J. Guid. Control Dyn.* **19**(6), 1201–1209 (1996)
- 32.113 J.K. Brock, R. Fuller, B. Kemper, D. Mleczko, J. Rodden, A. Tadros: GPS attitude determination and navigation flight experiment, *Proc. ION GPS* (1995) pp. 545–554
- 32.114 E.G. Lightsey, E. Ketchum, Th.W. Flatley, J.L. Crassidis, D. Freesland, K. Reiss, D. Young: Flight results of GPS based attitude control on the REX II spacecraft, *Proc. ION GPS* (1996) pp. 1037–1046
- 32.115 M. Unwin, P. Purivigraipong, A. da-Silva Curiel, M. Sweeting: Stand-alone spacecraft attitude determination using real flight GPS data from UOSAT-12, *Acta Astronaut.* **51**(1), 261–268 (2002)
- 32.116 S.M. Duncan, M.S. Hodgart, M.J. Unwin, R. Hebdon: In-orbit results from a space-borne GPS attitude experiment, *Proc. ION GNSS* (2007) pp. 2412–2423
- 32.117 S.F. Gomez: Attitude determination and attitude dilution of precision (ADOP) results for international space station global positioning system (GPS) receiver, *Proc. ION GPS* (2000) pp. 1995–2002
- 32.118 O. Montenbruck, M. Markgraf: Global positioning system sensor with instantaneous-impact-point reduction for sounding rockets, *J. Spacecr. Rockets* **41**(4), 644–650 (2004)
- 32.119 S.P. Powell, E.M. Klatt, P.M. Kintner: Plasma wave interferometry using GPS positioning and timing on a formation of three sub-orbital payloads, *Proc. ION GPS* (2002) pp. 145–154
- 32.120 B. Bull, J. Diehl, O. Montenbruck, M. Markgraf: Flight performance evaluation of three GPS receivers for sounding rocket tracking, *Proc. ION NTM* (2002) pp. 614–621
- 32.121 R. Broquet, N. Perrimon, B. Polle, P. Hyounet, P.A. Krauss, R. Draï, T. Voirin, V. Fernandez: Hi-NAV inertial / GNSS hybrid navigation system for launchers and re-entry vehicles, *Proc. NAVITEC, Noordwijk* (2010) pp. 1–6
- 32.122 P. Delaux, L. Bouaziz: Navigation algorithm of the atmospheric re-entry demonstrator, *Proc. AIAA 96-3754, AIAA Guidance, Navigation and Control Conf.*, San Diego (AIAA, Reston 1996), 1996
- 32.123 S.R. Steffes, S. Theil, M.A. Samaan, M. Conradt: Flight results from the SHEFEX2 hybrid navigation system experiment, *Proc. AIAA 2012-4991*,

AIAA Guidance, Navigation, and Control Conference, Minneapolis (AIAA, Reston 2012) pp. 1–14

32.124 M. Martin-Neira: A passive reflectometry and interferometry system (PARIS): Application to ocean altimetry, *ESA J.* **17**, 331–355 (1993)

32.125 R.A. Anthes: Exploring earth's atmosphere with radio occultation: Contributions to weather, climate and space weather, *Atmos. Meas. Tech.* **4**, 1077–1103 (2011)

32.126 S. Gleason, M. Adjrad, M. Unwin: Sensing ocean, ice and land reflected signals from space: Results from the UK-DMC GPS reflectometry experiment, *Proc. ION GNSS* (2005) pp. 1679–1685

Part F

Surveying

Part F Surveying, Geodesy and Geodynamics

33 The International GNSS Service

Gary Johnston, Symonston, Australia
Anna Riddell, Symonston, Australia
Grant Hausler, Symonston, Australia

34 Orbit and Clock Product Generation

Jan P. Weiss, Boulder, USA
Peter Steigenberger, Wessling, Germany
Tim Springer, Seeheim-Jugenheim,
Germany

35 Surveying

Chris Rizos, Kensington, Australia

36 Geodesy

Zuheir Altamimi, Paris, France
Richard Gross, Pasadena, USA

37 Geodynamics

Jeff Freymueller, Fairbanks, USA

33. The International GNSS Service

Gary Johnston, Anna Riddell, Grant Hausler

The International global navigation satellite system (GNSS) Service (IGS) is an organization devoted to the generation of high-precision GNSS data and products; a service that benefits science and society. It is a voluntary federation of over 200 self-funding agencies, universities, and research institutions in more than 100 countries. Established in 1992 and formally launched on 1st January 1994, the IGS has delivered an uninterrupted time series of products that are utilized by a broad spectrum of users. IGS products have evolved over time, including the provision of GNSS data for constellations other than GPS, and the addition of real-time GNSS data and products.

This chapter provides an overview of the IGS, including a brief history and details of the current organization and its key components. The various products offered by the IGS are described and an outlook of future activities is given.

33.1	Mission and Organization	967
33.1.1	Mission	967
33.1.2	Structure	968
33.2	Components	969
33.2.1	IGS Governing Board and Executive Committee	969
33.2.2	IGS Central Bureau	969
33.2.3	IGS Network	970
33.2.4	Analysis Centers	970
33.2.5	Data Centers (DCs)	970
33.2.6	Working Groups	971
33.3	IGS Products	972
33.3.1	Orbits and Clocks	972
33.3.2	Earth Orientation and Site Coordinates	973
33.3.3	Atmospheric Parameters	974
33.3.4	Biases	975
33.4	Pilot Projects and Experiments	976
33.4.1	Real-Time	976
33.4.2	Multi-GNSS	978
33.5	Outlook	981
	References	981

33.1 Mission and Organization

33.1.1 Mission

The stated mission of the IGS is as follows [33.1]:

The International GNSS Service provides, on an openly available basis, the highest-quality GNSS data, products, and services in support of the terrestrial reference frame; Earth observation and research; Positioning, Navigation and Timing (PNT); and other applications that benefit the scientific community and society.

The IGS promotes a culture of shared expertise to encourage global best practice for developing and delivering GNSS data and products worldwide. This collaborative approach encourages input from a diverse

user community to strengthen uptake and promote innovation. The IGS has a strong interface with providers of GNSS equipment and services, and the owners of GNSS themselves. Drawing on a global research community, the IGS ensures that new technologies and systems can be integrated into its routine products. To support best practice among the global user community, the IGS also develops and publicly releases standards, guidelines, and conventions relating to the collection and use of GNSS data and products.

The IGS is a key element of the Global Geodetic Observing System (GGOS [33.2]) and fulfils three essential roles. The first is to provide the global linkage between the other elements of the global geodetic observing network, namely Satellite Laser Ranging (SLR) systems, Very Long Baseline Interferometry (VLBI)

Table 33.1 Key dates and milestones for the IGS

Date	Event
Aug 1989	First ideas for an International GPS Service were presented at the IAG General Meeting in Edinburgh
Jun 1992	Start of 1992 IGS Test Campaign (ended 23 Sep 1992)
Aug 1993	IAG Approval for IGS at IAG Scientific Meeting in Beijing.
Jan 1994	Start of official IGS
May 2000	Selective Availability removed from GPS
Mar 2001	GLONASS Service Pilot Project commenced
Mar 2001	TIGA (GPS Tide Gauge Benchmark Monitoring) project established
Apr 2003	Ionosphere maps (IONEX) etc. became official IGS product
May 2003	First operational combined GPS/GLONASS analysis products released
Mar 2005	IGS renamed as International GNSS Service
Dec 2005	International Committee on GNSS created by the UN office of Outer Space Affairs
Aug 2011	Multi-GNSS Experiment (MGEX) Call for Participation
Jul 2013	Real-Time Pilot Project commenced

telescopes, and Doppler Orbitography and Radio positioning Integrated by Satellite (DORIS) ground beacons. These links are fundamental to generating the International Terrestrial Reference Frame (ITRF) [33.3]. Given the high cost of establishing and maintaining SLR and VLBI facilities, and the fact that direct collocation of SLR, VLBI, and DORIS is not always viable, IGS products provide a cost-effective means of geometrically linking these other observing techniques. The second role is to densify and improve the geometric distribution of the global geodetic network, allowing accurate modeling of satellite orbits and clocks, atmospheric behavior, and Earth processes like neo-tectonics. The final role is to provide the user segment with access to the ITRF, which is increasingly important as both the accuracy of publicly accessible GNSS positioning improves and the consequent need to better understand the relationship between the ITRF and national datums emerges.

There are numerous benefits for an organization or station operator to support the IGS, including increased accuracy of the reference frame in the organization’s region of interest; simpler and more accurate connections to the reference frame; increased accuracy of global positioning products, such as satellite orbit and clock products; and more accurate determination of transformations between realizations of the reference frame in that region and the respective national datum.

The adoption of a United Nations General Assembly Resolution in 2015 supporting the Global Geodetic Reference Frame is in many ways recognition of the role the IGS plays in supporting society and more specifically the UN Sustainable Development Goals [33.4].

The IGS strives to maintain an international federation with committed contributions from its members. It does this by providing effective leadership, management, and governance. While the value proposition for participation in the IGS varies for different contributors, its data and products are largely driven by user needs. IGS governance encourages an inclusive culture, and particular attention is given to outreach across regions and countries with lower than expected participation levels. The IGS also provides support and leadership to a multitude of other science programs. Considerable attention is paid to providing advocacy for the IGS into the Group on Earth Observations (GEO); the Global Earth Observation System of Systems (GEOSS); the Committee on Earth Observation Satellites (CEOS); and a variety of United Nations committees.

A comprehensive history of the IGS can be found in [33.5] and Table 33.1 lists key dates and milestones since the IGS was first conceived.

33.1.2 Structure

An overview of the IGS organization as of 2015 is provided in Fig. 33.1. Key elements include:

- The *Governing Board (GB)* that oversees the IGS activities establishes its policies and decides on the establishment of new activities and products.
- The *Central Bureau (CB)* that performs the overall coordination and day-to-day management of IGS activities.
- The *IGS Network* of globally distributed monitoring stations that provide continuous observations of all GNSS constellations.
- The *Analysis Centers (ACs)* that generate quality controlled products such as precise orbit and clock solutions, tropospheric and ionosphere maps, and station position estimates from GNSS observations.
- The *Data Centers (DCs)* that make all data and products available to the community.
- Various *Working Groups (WGs)* that provide technical guidance and expertise in specific fields to advance the product generation and to establish new data and processing standards.

The tasks of each organizational element are further described in Sect. 33.2.

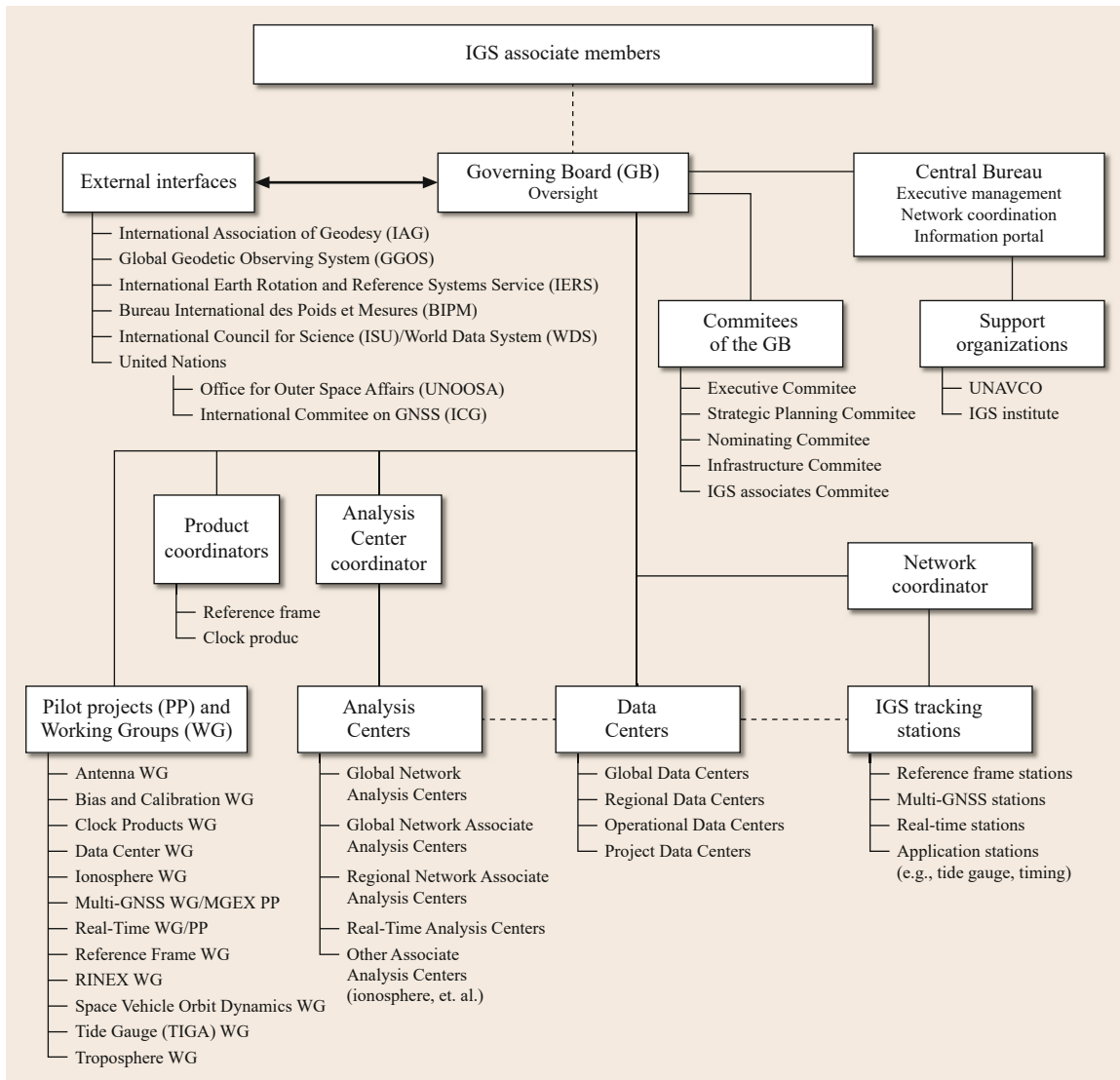


Fig. 33.1 Organizational structure of the IGS

33.2 Components

Key components of the IGS were identified in Fig. 33.1. This section provides a brief summary of the roles and responsibilities associated with each of these components. Further details are provided in the IGS Terms of Reference [33.1].

33.2.1 IGS Governing Board and Executive Committee

The Governing Board (GB) is the international body that sets policy for the IGS and exercises broad oversight of all IGS functions and components. It controls

the general activities of the IGS and implements restructuring if needed to increase the organization's efficiency and reliability for integrating and making full use of all available GNSS technologies.

The GB has an Executive Committee (EC) with specific responsibilities that allow it to act on behalf of the GB outside formal GB meetings.

33.2.2 IGS Central Bureau

The Central Bureau (CB) coordinates the day-to-day operations of the IGS. It is the executive arm of the

IGS GB with responsibilities for coordinating general aspects of IGS network operations; promoting compliance to IGS standards; monitoring network operations and providing quality assurance of data; maintaining documentation; organizing meetings and workshops; and coordinating development and publishing of IGS reports. The performance of the CB is formally reviewed by the GB at least every 5 years to ensure it is capable of fulfilling its long-term coordination role.

33.2.3 IGS Network

The foundation of the IGS is a global network of over 450 permanent and continuously operating stations of geodetic quality, which track signals from GPS. Increasingly signals from GLONASS, Galileo, BeiDou, QZSS, and several Space-Based Augmentation Systems (SBAS) are also tracked. To ensure continuous tracking of high-accuracy GNSS data, stations in the IGS network are developed to a minimum set of physical and operational standards defined by the *IGS Site Guidelines* [33.6]. For example, IGS infrastructure must be physically stable to support long-term operation of the IGS network, and any changes to a station's configuration should be carefully planned and documented to minimize discontinuities in the station's position time-series. Minimum scheduling requirements are also implemented to ensure that data from each station is transmitted as rapidly as possible to global and regional DCs for archiving and analysis.

Station locations in the global IGS tracking network are illustrated in Fig. 33.2.

33.2.4 Analysis Centers

ACs receive and process tracking data from one or more DCs for the purpose of producing IGS products. ACs are recognized by the GB as those groups which commit to deliver some or all of the core IGS products

(Sect. 33.3), within a specified time period, using designated IGS standards and conventions. Core products typically include satellite ephemerides, Earth rotation parameters, station coordinates, and clock information. The products are produced in Ultra-rapid, Rapid, Final, and Reprocessed versions for each AC.

Associate ACs (AACs) are a second category of AC that produce specialized products, including ionospheric information, tropospheric parameters, or station coordinates and velocities for global or regional subnetworks. Regional Network Associate Analysis Centers (RNAACs) and Global Network Associate Analysis Centers (GNAACs) are currently recognized by the GB. The functions of AACs continue to evolve as new capabilities and products emerge within the IGS.

Finally, the Analysis Center Coordinator (ACC) is responsible for combining products from each AC into a single set of orbit and clock products, which are the official IGS products made available to users through the Global DCs. The ACC monitors and assists the activities of ACs to ensure IGS objectives and standards for quality control, performance evaluation, and analysis are carried out. The ACC is a voting member of the IGS GB and interacts regularly with the CB and International Earth Rotation and Reference Systems Service (IERS). The responsibilities for the ACC typically rotate around the ACs with appointments and terms specified by the GB.

33.2.5 Data Centers (DCs)

The *Charter for IGS DCs* [33.7] defines three categories – Operational, Regional, and Global DCs – each of which builds redundancy into the IGS network. DCs are approved by the GB based on recommendations of the DCWG, and a demonstrated commitment to IGS principles and standards.

Operational DCs are in direct contact with IGS tracking sites. Their tasks include station monitoring,

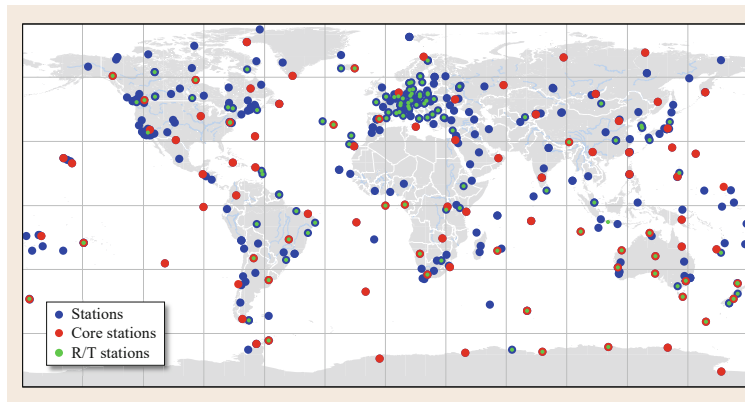


Fig. 33.2 Sites in the global IGS tracking network. Out of an overall set of about 470 stations available in Oct. 2015, the 90 core stations marked in red are used to establish the IGB08 reference frame. Green dots indicate stations with real-time data transmission capability

data validation, data formatting and exchange (e.g., [RINEX](#)), data compression, local archiving of GNSS data, and the electronic transmission of data to Regional and Global DCs [33.8]. Download schedules and data continuity requirements are specified in the IGS Site Guidelines for Operational DCs.

Regional DCs collect tracking data in the required exchange format from several Operational Centers and/or stations. They maintain a local archive of this data, provide on-line access to the data, and transmit the data from a subset of their sites (minimally, the IGS reference frame stations) to Global DCs. The stations managed by Regional DCs can be those of an individual agency or those located across a specific geographic region (Europe, Australia, etc.).

Global DCs are the main interfaces to the ACs and general user community. They receive, retrieve, archive, and provide online access to tracking data from Operational and Regional DCs. They are responsible for archiving and backing up IGS data and products, and exchanging data between other DCs to balance data holdings across the IGS network. At a minimum, Global DCs must archive GNSS data that has been sampled at 30 s intervals from IGS reference frame sites. As of 2015, the IGS comprises four global DCs hosted by institutions in the United States, France, and Korea:

- Crustal Dynamics Data Information System (CD-DIS [33.9])
- Institut National de l'Information Géographique et Forestière (IGN)
- Scripps Institution of Oceanography (SIO)
- Korean Astronomy and Space Science Institute (KASI).

Routine quality control is encouraged by all DCs to validate data prior to transmission.

33.2.6 Working Groups

The IGS has a number of working groups that focus on different aspects of product generation. These WGs also support IGS pilot projects (Sect. 33.4) to investigate future GNSS developments that could lead to the generation of new IGS products.

The current WGs are briefly summarized below along with their goals and objectives:

- *Antenna Working Group (AWG)* – To increase the accuracy and consistency of IGS products the AWG coordinates research on GNSS receiver and satellite antenna phase center determination, and manages official IGS antenna files and their formats.

- *Bias and Calibration Working Group (BCWG)* – Different GNSS observables are subject to different satellite biases that can degrade the IGS products. The BCWG coordinates research for retrieving and monitoring GNSS biases, and develops rules for handling these biases.
- *Clock Products Working Group (CPWG)* – The CPWG is responsible for aligning the combined IGS products to a highly precise timescale traceable to the world standard; Coordinated Universal Time (UTC).
- *Data Center Working Group (DCWG)* – The DCWG works to improve the provision of data and products from the Operational, Regional, and Global DCs, and recommends new DCs to the GB.
- *Ionosphere Working Group (IWG)* – The IWG produces global ionosphere maps of Ionosphere Vertical Total Electron Content (TEC). A major task of IWG is to make available global ionosphere maps from the TEC maps produced independently by Ionosphere Associate Analysis Centers (IAACs) within the IGS.
- *Multi-GNSS Working Group (MGWG)* – The MGWG supports the MGEX Project by facilitating estimation of intersystem biases and comparing the performance of multi-GNSS equipment and processing software. The MGEX Project was established to track, collate, and analyze all available GNSS signals including those from BeiDou, Galileo, and QZSS in addition to GPS and GLONASS satellites.
- *Reference Frame Working Group (RFGW)* – The RFGW combines solutions from the IGS ACs to form the IGS station positions and velocity products, and Earth rotation parameters for inclusion in the IGS realization of ITRF.
- *Real-Time Working Group (RTWG)* – The RTWG supports the development and integration of real-time technologies, standards, and infrastructure to produce high-accuracy IGS products in real time. The RTWG operates the IGS Real-Time Service (RTS) to support Precise Point Positioning (PPP) at global scales, in real time.
- *RINEX Working Group (RINEX-WG)* – The RINEX-WG jointly manages the RINEX format with the Radio Technical Commission for Maritime services-Special Committee 104 (RTCM-SC104). RINEX has been widely adopted as an industry standard for archiving and exchanging GNSS observations, and newer versions support multiple GNSS constellations.
- *Space Vehicle Orbit Dynamics Working Group (SVODWG)* – The SVODWG brings together IGS groups working on orbit dynamics and attitude

modelling of spacecraft. This work includes the development of force and attitude models for new GNSS constellations to fully exploit all new signals with the highest possible accuracy.

- *Tide Gauge (TIGA) Working Group* – TIGA is a pilot study for establishing a service to analyze GPS data from stations at or near tide gauges in the IGS network to support accurate measurement of sea-level change across the globe.

- *Troposphere Working Group (TWG)* – The TWG supports development of the IGS troposphere products by combining troposphere solutions from individual ACs to improve the accuracy of PPP solutions.

Information on the WG charters and membership can be found at [33.10]. WG chairs report to the IGS GB on a regular basis.

33.3 IGS Products

The primary objective of the IGS is to provide the reference GNSS products and observations for a wide variety of scientific and engineering users involving GNSS. To fulfil this role, the IGS produces a number of fundamental products such as:

- GNSS orbits and clocks
- Earth orientation parameters and station coordinates
- Ionosphere and troposphere parameters and
- Systematic bias estimates.

These high-quality products are used to support scientific applications such as the realization of the ITRF, monitoring the deformation of the solid Earth due to ocean tides and hydrology, and monitoring of sea-level change and associated climate change events. Increasingly the products are also used for sounding the atmosphere and producing ionospheric and tropospheric

maps. Lastly, the IGS products are used extensively to support precise positioning applications for industry and society.

33.3.1 Orbits and Clocks

In order for a user to achieve precise positioning (Chap. 25), knowledge of the orbits and clocks of the GNSS satellites is fundamental. The positioning accuracy is directly affected by errors in the satellite orbits and clocks. Relatively low accuracy orbit and clock information are transmitted through the GNSS navigation messages. Other more precise orbits and clock information are provided by the IGS and its individual ACs. An indicative list of the IGS orbit and clock products, together with their latency and availability, is provided in Tables 33.2 and 33.3.

Table 33.2 Accuracy, latency, continuity, availability, and sampling intervals for IGS orbit and clock products relating to GPS satellite orbits and satellite (sat) and station (stn) clocks as of 2013 (after [33.11, 12]). For definition of latency, continuity, and availability see [33.11]

GPS satellite ephemerides Satellite and station clocks		Sample interval	Accuracy	Latency	Continuity	Availability (%)
Broadcast (for comparison)	Orbits Sat. clocks	–	≈ 100 cm ≈ 5 ns RMS, 2.5 ns σ	Real time	Continuous	99.99
Ultra-rapid (predicted half)	Orbits Sat. clocks	15 min	≈ 5 cm ≈ 3 ns RMS, ≈ 1.5 ns σ	Predicted	4× daily, at 3 h, 9 h, 15 h, 21 h UTC	95
Ultra-rapid (observed half)	Orbits Sat. clocks	15 min	≈ 3 cm ≈ 150 ps RMS, ≈ 50 ps σ	3–9 h	4× daily, at 3 h, 9 h, 15 h, 21 h UTC	
Rapid	Orbits, Sat. & stn. clocks	15 min 5 min	≈ 2.5 cm ≈ 75 ps RMS, ≈ 25 ps σ	17–41 h	daily, at 17 h UTC	95
Final	Orbits, Sat. & stn. clocks	15 min 30 s (Sat) 5 min (Stn)	≈ 2 cm 75 ps RMS, 20 ps σ	12–18 d	weekly, Thursday	99
Real-time	Orbits Sat. clocks	5–60 s	≈ 5 cm 300 ps RMS, 120 ps σ	25 s 5 s	Continuous	95

Table 33.3 Accuracy, latency, continuity, availability, and sampling intervals for IGS orbit and clock products relating to GLONASS satellite ephemerides as of 2013 (after [33.11, 12])

GLONASS satellite orbits	Sample interval	Accuracy	Latency	Continuity	Availability (%)
Final	15 min	≈ 3 cm	12–18 d	Weekly, every Thursday	99

The IGS continuously monitors the performance of its products through its ACC. The ACC is in charge of monitoring the performance of the orbits and clocks by comparing the final products against each individual AC product. An example of this comparison is illustrated in Fig. 33.3 where each individual AC orbit is compared to the IGS final orbits via three-dimensional (3-D) differences.

33.3.2 Earth Orientation and Site Coordinates

The Earth Orientation Parameters (EOPs) are another product derived from the complex computations performed by the IGS (Table 33.4). EOPs specify the motion of the Earth's rotation pole and its irregularities through time. These parameters provide the connection or tie of the International Celestial Reference Frame (ICRF) to the ITRF and the IGS Terrestrial Reference Frame, known as the Igb. They consist of:

- The Universal Time (UT), which is the time of the Earth clock
- The length-of-day (LOD), which describes any excess in the revolution time
- The polar motion, which describes the varying position of the Earth's pole through coordinates x and y of the Celestial Ephemeris Pole (CEP) relative to the IERS reference pole
- The polar motion rate, which represents the velocity of the polar coordinates and how they vary through time.

The physical meaning of the individual parameters is further discussed in Chap. 2 of this Handbook and the IERS Conventions [33.14].

Site Coordinates

Site positions and velocities (Table 33.5) for the IGS stations are generated by the analysis centers on a weekly basis. The individual solutions are provided

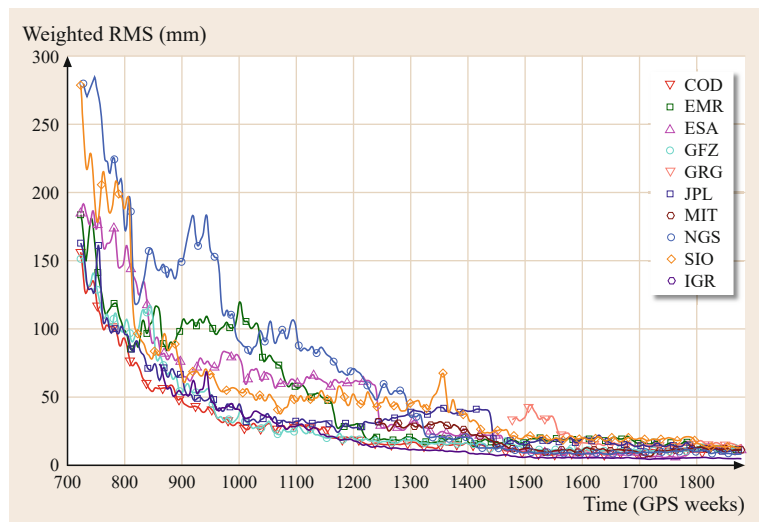


Fig. 33.3 Weighted RMS (mm) of the individual AC orbit solutions with respect to the IGS Final orbits for the period 1994 to Dec. 2015 (smoothed). Individual Analysis Centers are identified by their three-letter acronyms (COD: Center for Orbit Determination in Europe, Switzerland; EMR: Natural Resources Canada; ESA: European Space Agency; GFZ: GeoForschungsZentrum Potsdam, Germany; GRG: Groupe de Recherche de Géodésie Spatiale – Centre National d'Etudes Spatiales (CNES) and Collecte Localisation Satellites (CLS); JPL: Jet Propulsion Laboratory, USA; MIT: Massachusetts Institute of Technology, USA; NGS: National Geodetic Survey, National Oceanic and Atmospheric Administration (NOAA), USA; SIO: Scripps Institute of Oceanography, USA [33.13]). Image courtesy of Geoscience Australia and MIT

Table 33.4 IGS Earth orientation products (after [33.11, 12])

Earth rotation parameters		Sample interval	Accuracy	Latency	Continuity	Availability (%)
Ultra-rapid (predicted half)	Polar motion	Daily integrations	$\approx 200 \mu\text{s}$	Real time	4× daily, at 3 h, 9 h, 15 h, 21 h UTC	99
	Polar motion rate	at 0 h, 6 h, 12 h, 18 h UTC	$\approx 300 \mu\text{s/d}$			
	Length-of-day		$\approx 50 \mu\text{s}$			
Ultra-rapid (observed half)	Polar motion	Daily integrations	$\approx 50 \mu\text{s}$	3–9 h	4× daily, at 3 h, 9 h, 15 h, 21 h UTC	99
	Polar motion rate	at 0 h, 6 h, 12 h, 18 h UTC	$\approx 250 \mu\text{s/d}$			
	Length-of-day		$\approx 10 \mu\text{s}$			
Rapid	Polar motion	Daily integrations	$\approx 40 \mu\text{s}$	17–41 h	Daily, at 17h UTC	99
	Polar motion rate	at 12 h UTC	$\approx 200 \mu\text{s/d}$			
	Length-of-day		$\approx 10 \mu\text{s}$			
Final	Polar motion	Daily integrations	$\approx 30 \mu\text{s}$	11–17 d	Weekly, Wednesday	99
	Polar motion rate	at 12 h UTC	$\approx 150 \mu\text{s/d}$			
	Length-of-day		$\approx 10 \mu\text{s}$			

Note 1: $100 \mu\text{s} = 3.1 \text{ mm}$ and $10 \mu\text{s} = 4.6 \text{ mm}$ of equatorial rotation at the Earth's surface.

Note 2: The IGS uses VLBI from IERS Bulletin A to partially calibrate for LOD biases over a 21-day sliding window, but residual time-correlated LOD errors remain.

Table 33.5 IGS station coordinate products (after [33.11, 12])

Geocentric coordinates of IGS tracking stations (> 250 sites)		Sample interval	Accuracy	Latency	Continuity	Availability (%)
Final positions	Horizontal	Weekly	3 mm	11–17 d	Weekly, Wednesday	99
	Vertical		6 mm			
Final velocities	Horizontal	Weekly	2 mm/yr	11–17 d	Weekly, Wednesday	99
	Vertical		3 mm/yr			

in the Solution Independent Exchange (SINEX) format (Annex A.2.3), which facilitates combinations of different ACs but also of GNSS-derived solutions with other space-geodetic techniques.

Through its GNSS network and processing the IGS contributes to, extends, and densifies the ITRF. The ITRF provides an accurate and consistent frame, or datum, for referencing positions at different times and in different locations around the world. The IGS realization of ITRF, which extends the number of stations significantly, makes the reference frame easily accessible.

The IGS network shown in Fig. 33.2 includes a well-distributed subset (the *reference frame stations* indicated in red) called the IGB08 core network. This subnetwork is recommended for comparison or alignment of global solutions to the IGB08 reference frame, in order to mitigate the aliasing of station nonlinear motions into transformation parameters (network effect). It is used for the alignment of the IGS weekly combined solutions to IGB08.

IGS reference frame stations are the highest-quality GNSS stations in the world. This quality directly impacts the level of accuracy that can be achieved by using the ITRF. Requirements for stations include: a high-quality monument on stable crustal bedrock with excellent sky visibility; a long observing history; high-

quality, consistent, continuous, and complete raw data; minimal changes to equipment and its surroundings; and a commitment to keep the station operating for as long as possible. Full requirements are detailed in the IGS Site Guidelines [33.6]. These site requirements are stringent in order to ensure reliable measurements uniformly across the global network in support of projects such as sea-level change, which occurs at the millimeter level. Limitations in a reference frame negatively impact the accuracy of numerous scientific and positioning applications, especially in the region immediately surrounding the station.

33.3.3 Atmospheric Parameters

Other IGS products are developed by AACs in relation with WGs and pilot projects. These products include Zenith Troposphere Delay (ZTD; also known as Zenith Path Delay, ZPD) parameters and ionospheric vertical total electron content (VTEC) maps, which have application in climate and atmospheric research (Chaps. 38 and 39). A summary of the IGS atmospheric products is given in Table 33.6.

Troposphere

The ZTD products are generated from the IGS ACs through the processing of ground-based GNSS

Table 33.6 IGS atmospheric products (after [33.11, 12])

Atmospheric parameters	Sample interval	Accuracy	Latency	Continuity	Availability (%)
IGS final tropospheric delay (ZTD and gradients)	5 min	≈ 4 mm for ZTD	≈ 3 weeks	Daily	99
Ionosphere TEC grid	2 h, $5^\circ \times 2.5^\circ$ (lon./lat.)	2–8 TECU	≈ 11 d	Weekly	99
Rapid ionosphere TEC grid	2 h, $5^\circ \times 2.5^\circ$ (lon./lat.)	2–9 TECU	< 24 h	Daily	95

data [33.15, 16]. For the extraction of the ZTD and its horizontal gradients, observations of the surface pressure and temperature at the GNSS sites are needed. In order to produce these ZTD products, the IGS ACs need to use all of the aforementioned products as known parameters, that is orbits, clocks, and EOPs.

Ionosphere

The ionosphere products, which comprise a set of VTEC maps [33.17], are also a product of a GNSS processing strategy using orbits, clocks, and EOPs, which are derived from dual frequency observations. Another by-product of this is the estimates of the differential code biases (DCBs) (discussed in the following section). These ionosphere products are available from the rapid solutions with a latency of less than 24 hours, a final solution with a latency of approximately 11 days, and a predicted solution available both 1 and 2 days prior. Ionosphere products are provided in IONEX (Ionosphere Exchange) format (Annex A.2.4).

33.3.4 Biases

All GNSS observations are biased and contain unknown quantities characterized as systematic errors. The IGS, through the Bias and Calibration Working Group (BCWG), coordinates the production, research, and monitoring of these biases. The BCWG is responsible for defining the rules and procedures, which dictate the consistent handling and processing of these biases in an inhomogeneous GNSS environment.

The current set of IGS bias estimates comprises DCBs for GPS and GLONASS signals on the L1 and L2 frequencies:

- L1 P(Y)- or P-code minus L1 C/A-code biases (termed *P1-C1* in accord with the heritage two-letter RINEX 2 observation designations) for the GPS and GLONASS constellation for a moving 30-day combination (considering the bias estimates of the latest 30 daily solutions).
- L2 P(Y)- or P-code minus L2C or L2 C/A-code biases (termed *P2-C2*) for the GPS and GLONASS constellations for a moving 30-day combination.

- P1-P2 bias values for GPS and GLONASS as a by-product of the ionospheric analysis as monthly values.

These biases are generated by two methods: the indirect and direct method. In the indirect estimation process the biases are generated as estimable parameters in the clock determination process. In the direct process, the biases are generated directly from the differences of the observations of the different signals.

In the face of GPS and GLONASS modernization programs and upcoming GNSS, like the European Galileo and the Chinese BeiDou, an increasing number of types of biases are expected [33.18]. Some of the future biases products that the IGS will have to produce in the context of multi-GNSS multifrequency signals are:

- Transition to RINEX 3.xx compatible bias types and designations (e.g., C1W-C1C, C2W-C2S, C1W-C2W, etc.)
- Biases for particular linear combinations such as wide-lane biases (WLB), ionosphere-free biases (LCB), narrow-lane biases (NLB)
- GLONASS interfrequency code biases
- GLONASS differential code-phase biases for ambiguity resolution
- Uncalibrated phase delays (UPD) relevant to undifferenced integer fixing for precise point positioning
- GPS quarter-cycle phase offset issues (specifically between L2W and L2C)
- Differentials code biases for new signals and constellations
- Absolute or observable code bias values, being consistent with respect to each DCB set and with respect to the signals currently used for the clock offset determination.

The generation of multi-GNSS DCB products has been initiated in the frame of the IGS multi-GNSS Experiment (MGEX; Sect. 33.4.2), where prototype DCB products covering new signals and constellations are made available by selected analysis centers since 2014 [33.19].

33.4 Pilot Projects and Experiments

Throughout its history, the IGS has organized various campaigns, experiments, and pilot projects to actively support emerging GNSS developments and to prepare for the generation of new IGS products. Past examples include the International GLONASS Experiment [33.20] and the subsequent International GLONASS Service (IGLOS) pilot project [33.21], which laid the foundation for today's IGS GLONASS products. Another example is the GPS Tide Gauge Benchmark Monitoring (TIGA) project, which applies precise GPS results to monitor vertical motion of tide gauges. The TIGA pilot project became the TIGA Working Group in 2010. More recent activities include the IGS RTS and the Multi-GNSS Experiment (MGEX), which are discussed in this section.

33.4.1 Real-Time

Launched in April 2013, the IGS RTS generates GNSS products in real time to support Precise Point Positioning (PPP; Chap. 25) at global scales. It makes use of raw data which are continuously streamed from a subset of high-quality GNSS receivers in the global IGS network. The RTS products consist of GNSS satellite orbit and clock corrections. Data and products from the RTS provide real-time access to the global reference frame. The RTS is a public service made openly available to users that hold a free subscription.

Prior to launching the RTS, access to the global reference frame using IGS products has been *ex post facto* or *after-the-fact*. The RTS makes these products available with little or no latency to support PPP in real time. This enables scientific, educational, and commercial applications at worldwide scales, including geophysical hazard detection and warnings, conventional and space weather forecasting, time synchronization, and performance monitoring of GNSS constellations [33.22].

The RTS is built on the network of tracking stations, DCs and Real-Time Analysis Centers (RTACs) that underpin the global IGS network. Planning for the RTS has been underway since 2002, with careful attention given to network design and management, algorithm development, product generation, and defining real-time protocols and standards for accessing data and products. Support is provided by over 120 station operators, multiple DCs, and 10 RTACs around the world.

A brief outline of the RTS network and its data and products is provided below. Up-to-date information on the network status, product types, and performances as well as user access is provided through the IGS RTS website [33.23].

RTS Network

The RTS is built on the global IGS infrastructure that functions as a world standard for high-precision GNSS data and products. Contributions to the IGS are made on a collaborative and best-efforts basis, meaning the RTS including all data and products is offered without a service guarantee. However, it is these global contributions that build redundancy into the RTS and its products. The global RTS architecture ensures that a reliable flow of data and products is available without interruption.

The RTS comprises over 120 globally distributed GNSS stations maintained by a wide variety of local and region IGS operators. These stations deliver 1 Hz data to real-time DCs within the IGS network, with typical latencies of 3 s or less. The distribution of real-time tracking stations within the overall IGS network is illustrated in Fig. 33.2.

Network redundancy and a global distribution of stations are needed to provide full coverage and a reliable flow of real-time data. Coverage is challenging in some areas, particularly in regions of vast open ocean. Additional contributions to the network are encouraged where possible, however new stations in the RTS network must adhere to a minimum set of infrastructure standards and IGS best practices for real-time operations. Examples of these best practices are provided in [33.24] and illustrated in Fig. 33.4:

- Real-time data should be transmitted to a minimum of two separate real-time DCs.
- Stations that contribute to the realization of the IGS reference frame should be operated in real time to

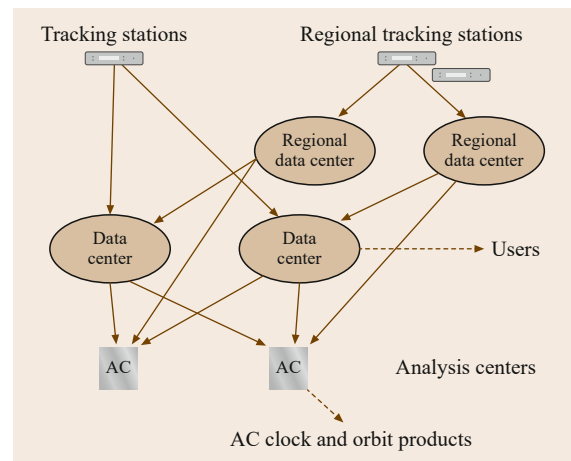


Fig. 33.4 Real-time data is streamed to multiple DCs and ACs to build redundancy into the RTS (after [33.24])

guarantee reliable alignment of the real-time products to a stable reference frame.

- RTACs are encouraged to ingest data from two or more global data centers.

Individual correction products are produced within each RTAC once the real-time data streams have been received from each DC. The final orbit and clock correction products delivered from the RTS are actually a combination of the individual RTAC correction products, producing a more reliable and stable set of products than any single RTAC product alone.

This highly redundant design is made possible through the contribution of 10 RTACs within the RTS. Operational responsibility for producing the official combination products lies with the IGS RTAC Coordinator (RTACC), which is currently contributed by European Space Agency (ESA/ESOC) in Darmstadt, Germany. Figure 33.5 illustrates the RTS architecture for combining and distributing RTAC solutions via multiple DCs worldwide.

To support ongoing development and management of the RTS, the IGS RTWG addresses issues pertaining to infrastructure management and data analysis. The activities of the RTWG consist of planning, designing, and implementing next stages for the RTS, including development of multi-GNSS correction products and associated standards, along with outreach to new IGS participants and users of the RTS. This work is guided by the broader IGS strategy to enhance standards and best practices for GNSS infrastructure management and data availability to benefit the global user community.

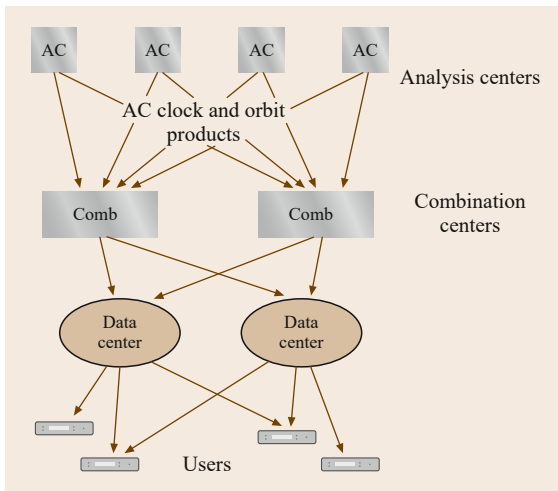


Fig. 33.5 Orbit and clock corrections produced by each AC in the RTS are combined to deliver a more reliable and stable correction product to users (after [33.24])

RTS Data and Products

To support global interoperability and integration of GNSS technologies and systems, the IGS develops and maintains standards and formats for disseminating GNSS data and products. IGS joined the Radio Technical Commission for Maritime Services Special Committee 104 (RTCM-SC104) in 2008 and adopted the RTCM-3 format for GPS and GLONASS observation messages shortly after. In Fig. 33.4, for example, the real-time GNSS data streams flowing from IGS tracking stations to the RTACs, via each DC, are formatted in the latest version RTCM-3 (most recently v3.2 [33.25]). Annex A.1.3 of this Handbook also describes the new RTCM3 Multi-Signal Message (MSM) format being developed to handle all GNSS constellations, signals, and observation types as part of the IGS Multi-GNSS Experiment (MGEX). Prototype MSM data streams are already being tested through the MGEX project and receiver manufacturers have started to release firmware supporting MSM [33.26].

The IGS has also adopted the RTCM-State Space Representation (RTCM-SSR [33.27]) format for disseminating real-time orbit and clock correction messages. RTCM-SSR currently supports GPS and GLONASS constellations. These orbit and clock corrections are expressed within the International Terrestrial Reference Frame 2008 (ITRF08) and are designed to enable real-time PPP. The combined resolution of the RTCM-SSR corrections supports millimeter-accuracy corrections, and these corrections are broadcast over the Internet using the Network Transport of RTCM by Internet Protocol (NTRIP, [33.28]). NTRIP is an RTCM standard for disseminating and receiving RTCM-SSR messages.

Official IGS products from the RTS service are described in the IGS Strategic Plan 2013-2016 and briefly summarized in Table 33.7. Aside from orbit and clock corrections, the RTS also provides real-time access to broadcast ephemeris through the two data streams described below:

- **RTCM3EPH:** Broadcast ephemeris data for GPS, GLONASS, and Galileo satellites. This data stream is derived from receivers in the real-time IGS global network and encoded in RTCM-3 messages. The complete set of messages is repeated every 5 s.
- **RTCM3EPH01:** A GPS-only broadcast ephemeris stream also derived from the real-time IGS global network and encoded in RTCM Version 3 messages with a 5 s repetition rate.

Along with the state-space correction messages, these broadcast ephemeris can be used to reconstruct

Table 33.7 Content description of the RTS Product Streams (IGS, 2015b). APC: Antenna Phase Center; CoM: Center-of-Mass (not part of current RTCM-SSR standard). The figures in brackets next to each RTCM message ID denote the message sample interval in seconds

Stream name	Description	Ref. point	RTCM messages	Provider/Sol. ID	Bandwidth (kbits/s)	Combination center
IGS01	Orbit/Clock correction, single-epoch combination	APC	1059 (5), 1060 (5)	258/1	1.8	ESA/ESOC
IGC01	Orbit/Clock correction, single-epoch combination	CoM	1059 (5), 1060 (5)	258/9	1.8	ESA/ESOC
IGS02	Orbit/Clock correction, Kalman filter combination	APC	1057 (60), 1058 (10), 1059 (10)	258/2	0.6	BKG
IGS03	Orbit/Clock correction, Kalman filter combination	APC	1057 (60), 1058(10), 1059(10), 1063(60), 1064(10), 1065(10)	258/3	0.8	BKG

RTCM message types:

1057 GPS orbit corrections to Broadcast Ephemeris
 1063 GLONASS orbit corrections to Broadcast Ephemeris
 1058 GPS clock corrections to Broadcast Ephemeris
 1064 GLONASS clock corrections to Broadcast Ephemeris
 1059 GPS code biases
 1065 GLONASS code biases
 1060 Combined orbit and clock corrections to GPS Broadcast Ephemeris.

the precise orbit and clock information for use in real time or near-real-time PPP applications.

Standard IGS data and products are traditionally made available to users after-the-fact once a sufficient period of GNSS observation and processing has been undertaken to accurately model satellite orbits and clocks. The latency of these corrections ranges from hours to days to weeks depending on the final accuracy requirement. For example, *Final* orbit and clock products provide the highest accuracy and are delivered with a latency of 12–18 days. By contrast, the current RTS architecture enables real-time orbit and clocks products for GPS to be produced with an average latency of 25 s. The final accuracy of the RTS products is less than that of the Final IGS products, but sufficient enough to support real-time PPP. The accuracy, latency, continuity, and availability of all IGS orbit and clock products are compared in Table 33.2 above for the year 2013. Up-to-date performance monitoring results of the real-time products are provided at the IGS RTS website [33.23] on a routine basis.

33.4.2 Multi-GNSS

Today there are many GNSS signals available from multiple satellite navigation systems in addition to the well-known US Global Positioning System (Table 33.8). Other global systems include the Russian Global'naya Navigatsionnaya Sputnikovaya Sistema (GLONASS), the Chinese BeiDou Navigation Satellite

System (BDS), and Europe's Galileo. In addition to the global systems, there also exist regional systems such as the Japanese Quasi-Zenith Satellite System (QZSS) and the Indian Regional Navigation Satellite System (IRNSS/NavIC) as well as various Satellite-Based Augmentation Systems. This multi-GNSS environment offers various advantages to the users:

- Increased signal availability, even in undesirable environments (such as urban canyons)
- Increased number of frequency bands to improve robustness against interference
- Superior continuity of service and reduced dependency on a single system
- Efficiency gains due to faster ambiguity resolution
- Better reliability and redundancy, which enhances outlier detection
- Improvements in accuracy of position estimates.

The IGS Multi-GNSS Working Group (MGWG) was formed by the IGS in recognition of the rapidly evolving GNSS landscape to explore and promote the use of new constellations and navigation signals. The core activity of the MGWG is the Multi-GNSS Experiment (MGEX), which aims to build up a network of sensor stations, characterize the space segment and user equipment, develop theory and data-processing tools, and generate data products for the emerging satellite systems. The MGWG works closely with other IGS entities, such as the Data Center Working Group, the

Table 33.8 Global and regional satellite system deployment status and transmitted signals as of Oct. 2016. Brackets denote satellites not yet declared operational

System	Type	Signals	Satellites
GPS	IIR	L1 C/A, L1/L2 P(Y)	12
	IIR-M	+L2C	7
	IIF	+L5	12
GLONASS	M	L1/L2 C/A + P	23
	M+	+L3	1
	K1	+L3	1+(1)
BeiDou-2	GEO	B1, B2, B3	5+(1)
	IGSO	B1, B2, B3	6
	MEO	B1, B2, B3	3
BeiDou-3	IGSO	B1, L1, B2, E5a/b/ab	(2)
	MEO	B1, L1, B2, E5a/b/ab	(3)
Galileo	IOV	E1, E6, E5a/b/ab	3+(1)
	FOC	E1, E6, E5a/b/ab	6+(4)
QZSS	IGSO	L1 C/A, L1C, L1 SAIF, L2C, E6 LEX, L5	1
	IGSO	L5, S	4
IRNSS/NavIC	IGSO	L5, S	4
	GEO	L5, S	3

Antenna Working Group, and the Infrastructure Committee to achieve these goals.

The Multi-GNSS EXperiment was launched in February 2012. It was initially targeted at the global tracking of new and modernized GNSS signals as well as the build-up of analysis center capabilities to process such data and to generate associated multi-GNSS products. Within a couple of years, most of these goals have been reached and the IGS can offer its user a global network with multi-GNSS tracking capabilities supported by the corresponding data centers, analysis centers, and prototype products [33.26].

MGEX Network

Starting from a handful of individual stations available in early 2012, the MGEX network has grown to almost 130 sensor stations providing global/regional coverage

of GPS, GLONASS, BeiDou, Galileo, and QZSS in late 2015. Figure 33.6 illustrates the location of stations offering tracking of at least one of the new constellations in addition to GPS or GLONASS.

The MGEX network contains a diverse assortment of receiver and antenna equipment with five basic receiver types and eight main antenna types. All of the user equipment in the MGEX network is recognized and characterized by the IGS in the equipment description file. Many of the stations contain multiple receivers connected to the same antenna, known as a *zero-baseline* setup, which provides a basis for cross-validation of equipment performance. Some of the sites also contain multiple stations for short baseline comparison experiments. The variety of equipment used in the tracking network is both an asset as well as a challenge. The diversity of deployed receivers and antennas poses a challenge for consistent data processing, but allows a greater understanding of the types of data received and how the assessment of navigation signals is to proceed. The cross comparison of different receivers can contribute directly to design improvement by receiver manufacturers.

During the initial deployment of the MGEX network, relaxed site requirements were imposed on new station contributions and the network was essentially operated independently and in parallel to the legacy IGS tracking network for GPS and GLONASS to avoid any adverse impact on the standard IGS product generation. With the ongoing modernization of the legacy stations and the improved quality of new MGEX contributions, the vast majority of MGEX stations could ultimately be integrated into the standard IGS network in the course of 2015. Users can now take advantage of a single network of stations meeting the high-quality standards imposed by the IGS site guidelines. While only a subset of all IGS stations is presently multi-GNSS capable, the fraction of such stations is expected to grow continuously over time in the years to come.

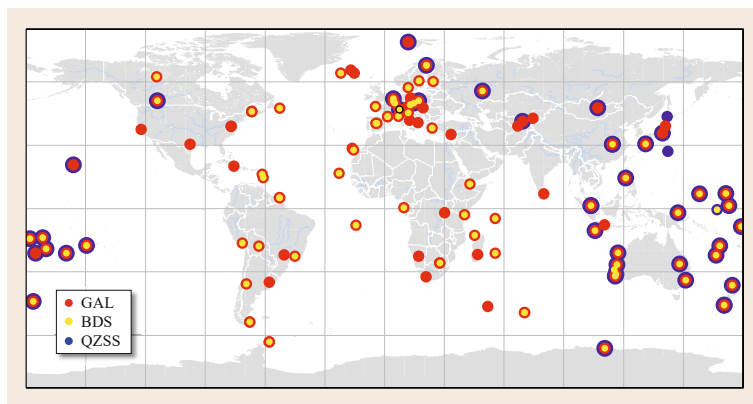


Fig. 33.6 MGEX network of stations (Oct. 2015)

Table 33.9 MGEX analysis centers generating precise multi-GNSS orbit and clock products on a routine basis (Oct. 2016)

Institution	Constellations
CNES/CLS, France [33.30]	GPS + GLO + GAL
CODE, Switzerland [33.31]	GPS + GLO + BDS + GAL + QZS
GFZ, Germany [33.32, 33]	GPS + GLO + BDS + GAL + QZS
JAXA, Japan	GPS + QZS
TUM, Germany [33.34]	GAL + QZS
Wuhan Univ., China [33.35]	GPS + GLO + BDS + GAL + QZS

All IGS multi-GNSS stations provide offline observation files in RINEX3 format [33.29], which supports all required observation types and constellations. Data archives are hosted by established IGS DCs including NASA's Crustal Dynamics Data Information System (CDDIS) in the United States, the French National Geographic Institute (IGN), and the German Federal Office for Cartography and Geodesy (BKG) Daily RINEX3 files with observation rates of 30 s are made available for all multi-GNSS stations and a subset of stations also delivers high-rate data files with 1 Hz observation data.

Next to the offline data, roughly 60% of all multi-GNSS stations also provide their data as real-time data streams. Where needed, manufacturer-specific data formats are encoded in the RTCM3 Multiple-Signal-Message (MSM) format once received from the individual sites and transferred to a dedicated MGEX NTRIP caster hosted by BKG in Frankfurt. The dedicated caster provides an experimental platform on which the new MSM format can be tested and facilitates user software adaptation.

MGEX Products

The data collected by the MGEX network provide the basis for the generation of precise orbit and clock products for the new constellations. As of late 2015, six analysis centers contribute such products to the MGEX project on a regular basis (Table 33.9). While early products were often confined to individual GNSSs, a growing number of ACs have lately moved to generating five-constellation products with a common underlying time and reference frame. Except for the Indian Regional Navigation Satellite System, the MGEX products thus cover all legacy and emerging navigation systems. Addition of IRNSS is foreseen, once an adequate number of monitoring stations with dual-frequency IRNSS tracking capability becomes available within the IGS.

Clock products were initially confined to 5 min or 15 min but have later been made available with sam-

pling intervals of down to 30 s. Furthermore, various ACs also provide associated data such as Earth orientation parameters, intersystem biases, or estimated station coordinates along with their orbit and clock products.

Other than for GPS and GLONASS, no combination process has yet been implemented within the IGS for precise orbit and clock products of the new constellations, but cross-comparison of different ACs as well as satellite laser ranging can serve to assess the precision or accuracy of the various products. For Galileo, a performance at the 10–20 cm level (3-D rms orbit consistency and differences w.r.t. SLR observations) has been demonstrated in [33.36] for the 2013–2014 time frame. Further improvements are expected through better characterization of the spacecraft and respective refinements of radiation pressure models or antenna phase center variations. A performance at the few dm-level is also achieved for BeiDou MEO and IGSO satellite, whereas the orbit consistency for geostationary satellites is limited to a few meters [33.37]. The degraded GEO orbit determination accuracy reflects the adverse impact of a near-static viewing geometry, which does not enable a reliable determination of the along-track position from one-way pseudorange and carrier-phase observations.

Even though the accuracy of the multi-GNSS products lags behind the performance of the standard IGS products for GPS and GLONASS, they pave the way for a full exploitation of new signals and constellations in navigation, surveying, geodesy, and remote sensing.

Special products provided in the frame of MGEX include combined multi-GNSS broadcast ephemeris data as well as multi-GNSS DCBs. The combined broadcast ephemeris data have initially served as a substitute for constellations not yet covered by the precise orbit and clock products, but are of continued interest because of their lower latency and the access to GNSS-specific system time scales.

An understanding of DCB is a prerequisite for processing multiconstellation code observations, which are essential in many navigation applications. They are also essential for many nonnavigation applications such as time transfer and ionospheric analysis for correction of pseudorange differences. The need for comprehensive DCB analysis is increasing as the rate of signals offered by new and modernized satellite systems rises. Prototype DCB products generated by the German Aerospace Center (DLR) and the Chinese Academy of Science (CAS) are available through MGEX at the CDDIS and IGN product archives. GPS, GLONASS, Galileo, and BeiDou DCBs are derived from pseudorange differences corrected for ionosphere path delays [33.19] and are provided in a preliminary version of the Bias SINEX format, which is currently under development within the

IGS and will serve as the new standard for the exchange of phase and code bias information.

For real-time users, early services include a combined multi-GNSS broadcast ephemeris stream (including GPS, GLONASS, Galileo, BeiDou, QZSS, and SBAS) prepared by BKG as well as Galileo orbit and clock corrections generated by the CNES/ILS analysis center.

33.5 Outlook

The IGS continues to evolve into a truly multi-GNSS service. It is expected that the new constellations of positioning satellites will be fully operational by 2020. This has driven the IGS to commence a strategic review of its activities, products, and services. While it is clear that the IGS has a role to play in producing products for all constellations (orbits and clocks), it is not so clear which combinations of signals will be utilized for the products that are nonconstellation-specific including the reference frame and atmospheric products. A considerable body of research is currently being undertaken to establish evidence to support such decisions.

As the uptake of GNSS as a public utility continues, an emerging need to understand the accuracy and integrity of all positioning satellites, including their system differences, is becoming evident. The use of GNSS now extends well beyond the military and science applications, including many industrial and Location Based Services (LBS) applications. The IGS has been cooper-

ating with Working Group A of the International Committee on Global Navigation Satellite Systems (ICG) to develop a common understanding of the requirement for system monitoring through the International GNSS Monitoring and Assessment (IGMA) subgroup.

Lastly, through GGOS, it has become clear that SLR observations to GNSS satellites and GNSS observations on non-GNSS satellites like GRACE, have a strong role to play in improving our understanding of observational errors, and therefore improving the accuracy of IGS products. The IGS will continue to collaborate with other elements of the GGOS to integrate these observations into the product generation.

For further information about MGEX, including news, constellation status, network and station information, data holdings, real-time data, and products, users are referred to the IGS MGEX website [33.38]. In the long run, it is planned to integrate all MGEX products into the regular IGS processing to offer a reliable and high-performance multi-GNSS service to the GNSS community.

Acknowledgments. The authors gratefully acknowledge valuable contributions to the chapter made by Stavros Melachronis. They would also like to acknowledge the support received from the IGS Central Bureau, the IGS working group chairs, and various contributing agencies.

References

- 33.1 International GNSS Service: *Terms of Reference* (IGS, Pasadena, 2014) <http://kb.igs.org/hc/en-us>
- 33.2 H.-P. Plag, M. Pearlman: *Global Geodetic Observing System: Meeting the Requirements of a Global Society on a Changing Planet in 2020* (Springer, Berlin 2009)
- 33.3 Z. Altamimi, X. Collilieux: IGS contribution to the ITRF, *J. Geod.* **83**(3/4), 375–383 (2009)
- 33.4 United Nations: A global geodetic reference frame for sustainable development, Resolution A/RES/69/266 adopted by the General Assembly on 26 Feb. 2015 (United Nations, New York 2015)
- 33.5 G. Beutler, A.W. Moore, I.I. Mueller: The International Global Navigation Satellite Systems Service (IGS): Development and achievements, *J. Geod.* **83**(3/4), 297–307 (2009)
- 33.6 IGS Central Bureau: *IGS Site Guidelines* (Infrastructure Committee, IGS Central Bureau, Pasadena 2015) <http://kb.igs.org/hc/en-us>
- 33.7 IGS: *Charter for IGS Data Centers – Definition of IGS Data Center Activities* (IGS, Pasadena 2010) <http://kb.igs.org/hc/en-us>
- 33.8 C. Noll, Y. Bock, H. Habrich, A. Moore: Development of data infrastructure to support scientific analysis for the International GNSS Service, *J. Geod.* **83**(3/4), 309–325 (2009)
- 33.9 C.E. Noll: The Crustal Dynamics Data Information System: A resource to support scientific analysis using space geodesy, *Adv. Space Res.* **45**(12), 1421–1440 (2010)
- 33.10 IGS: IGS Working Groups website, <http://igs.org/wg>
- 33.11 IGS: Strategic Plan 2013–2016 (IGS Central Bureau, Pasadena 2013) <http://kb.igs.org/hc/en-us>
- 33.12 IGS: IGS Products website, <http://www.igs.org/products/>
- 33.13 IGS: IGS Analysis Center Coordinator website <http://acc.igs.org/>

- 33.14 G. Petit, B. Luzum: *IERS Conventions (2010)*, IERS Technical Note No. 36 (Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt 2010)
- 33.15 S.H. Byun, Y.E. Bar-Sever: A new type of troposphere zenith path delay product of the international GNSS service, *J. Geod.* **83**(3/4), 1–7 (2009)
- 33.16 C. Hackman, G. Guerova, S. Byram, J. Dousa, U. Hugentobler: International GNSS Service (IGS) troposphere products and working group activities, FIG Work. Week 2015, Sofia (FIG, Copenhagen 2015) pp. 1–14
- 33.17 M. Hernández-Pajares, J.M. Juan, J. Sanz, R. Orus, A. García-Rigo, J. Felten, A. Komjathy, S.C. Schaer, A. Krankowski: The IGS VTEC maps: A reliable source of ionospheric information since 1998, *J. Geod.* **83**(3/4), 263–275 (2009)
- 33.18 S. Schaer: Biases and calibration working group technical report 2014. In: *IGS Technical Report*, ed. by Y. Jean, R. Dach (IGS Central Bureau, Pasadena 2014)
- 33.19 O. Montenbruck, A. Hauschild, P. Steigenberger: Differential code bias estimation using Multi-GNSS observations and global ionosphere maps, *Navigation* **61**(3), 191–201 (2014)
- 33.20 P. Willis, J. Slater, G. Beutler, W. Gurtner, C. Noll, R. Weber, R.E. Neilan, G. Hein: The IGEX-98-campaign: Highlights and perspective, *Geod. Beyond 2000*, Int. Assoc. Geod. Symp., Vol. 121, ed. by K.-P. Schwarz (Springer, Berlin 2000) pp. 22–25
- 33.21 R. Weber, J.A. Slater, E. Fagnier, V. Glotov, H. Habrich, I. Romero, S. Schaer: Precise GLONASS orbit determination within the IGS/IGLOS pilot project, *Adv. Space Res.* **36**(3), 369–375 (2005)
- 33.22 IGS: *IGS Real-Time Service Fact Sheet* (IGS, Pasadena 2014) <http://kb.igs.org/hc/en-us>
- 33.23 IGS: IGS Real-Time Service website <http://igs.org/rtss>
- 33.24 M. Caissy, L. Agrotis, G. Weber, M. Hernandez-Pajares, U. Hugentobler: Coming soon – The international GNSS real-time service, *GPS World* **23**(6), 52 (2012)
- 33.25 RTCM: RTCM Standard 10403.2 Differential GNSS Services, Version 3 with Amendment 2, 7 Nov. 2013 (RTCM, Arlington 2013)
- 33.26 O. Montenbruck, P. Steigenberger, R. Khachikyan, G. Weber, R.B. Langley, L. Mervart, U. Hugentobler: IGS-MGEX: Preparing the ground for multi-constellation GNSS science, *Inside GNSS* **9**(1), 42–49 (2014)
- 33.27 M. Schmitz: RTCM state space representation messages, status and plans, PPP-RTK Open Stand. Symp., Frankfurt (BKG, Frankfurt am Main 2012) pp. 1–31
- 33.28 G. Weber, D. Dettmering, H. Gebhard, R. Kalafus: Networked transport of RTCM via internet protocol (Ntrip) – IP-streaming for real-time GNSS applications, *Proc. ION GPS 2005*, Long Beach (ION, Virginia 2005) pp. 2243–2247
- 33.29 IGS: RINEX – The Receiver Independent Exchange Format – Version 3.03 14 Jul. 2015 (IGS RINEX WG and RTCM-SC104, 2015)
- 33.30 S. Loyer, F. Perosanz, F. Mercier, H. Capdeville: MGEX activities at CNES-CLS Analysis Centre, IGS Workshop 2012, Olsztyn (IGS, Pasadena 2012)
- 33.31 L. Prange, R. Dach, S. Lutz, S. Schaer, A. Jäggi: The CODE MGEX orbit and clock solution. In: *IGS 150 Years, IAG Symposia*, International Association of Geodesy Symposia, Vol. 143, ed. by C. Rizos, P. Willis (Springer, Berlin, Heidelberg 2015) pp. 767–773
- 33.32 M. Uhlemann, G. Gendt, M. Ramatschi, Z. Deng: GFZ global multi-GNSS network and data processing results. In: *IGS 150 Years, IAG Symposia*, International Association of Geodesy Symposia, Vol. 143, ed. by C. Rizos, P. Willis (Springer, Berlin, Heidelberg 2015) pp. 673–679
- 33.33 Z. Deng, M. Ge, M. Uhlemann, Q. Zhao: Precise orbit determination of BeiDou Satellites at GFZ, IGS Workshop 2014, Pasadena (IGS, Pasadena 2014)
- 33.34 P. Steigenberger, A. Hauschild, O. Montenbruck, C. Rodriguez-Solano, U. Hugentobler: Orbit and clock determination of QZSS-1 based on the CONGO network, *Navigation* **60**(1), 31–40 (2013)
- 33.35 J. Guo, X. Xu, Q. Zhao, J. Liu: Precise orbit determination for quad-constellation satellites at Wuhan University: Strategy, result validation, and comparison, *J. Geod.* **90**(2), 143–159 (2016)
- 33.36 P. Steigenberger, U. Hugentobler, S. Loyer, F. Perosanz, L. Prange, R. Dach, M. Uhlemann, G. Gendt, O. Montenbruck: Galileo orbit and clock quality of the IGS multi-GNSS experiment, *Adv. Space Res.* **55**(1), 269–281 (2015)
- 33.37 F. Guo, X. Li, X. Zhang, J. Wang: Assessment of precise orbit and clock products for Galileo, BeiDou, and QZSS from IGS Multi-GNSS Experiment (MGEX), *GPS Solutions* (2016), doi:[10.1007/s10291-016-0523-3](https://doi.org/10.1007/s10291-016-0523-3)
- 33.38 IGS: IGS Multi-GNSS Experiment (MGEX) website <http://igs.org/mgex>

34. Orbit and Clock Product Generation

Jan P. Weiss, Peter Steigenberger, Tim Springer

Many sophisticated Global Navigation Satellite System (GNSS) applications require high-precision satellite orbit and clock products. The GNSS orbits and clocks are usually derived from the analysis of tracking data collected by a globally distributed GNSS receiver network. The estimation process adjusts parameters for the satellite orbits, transmitter and receiver clocks, station positions, tropospheric delays, Earth orientation, intersystem and inter-frequency biases, and carrier-phase ambiguities. The estimation requires detailed modeling of geophysical processes, atmospheric and relativistic effects, receiver tracking modes, antenna phase centers, spacecraft properties, and attitude control algorithms. This chapter describes precise orbit and clock determination of the GNSS constellations as performed by the analysis centers of the International GNSS Service, including models, estimation strategies, products, and the combination of orbit and clock solutions.

34.1	Global Tracking Network	984
34.2	Models	985
34.2.1	Reference Frame Transformation	985
34.2.2	Site Displacement Effects	985
34.2.3	Tropospheric Delay	988
34.2.4	Ionospheric Delay	988
34.2.5	Relativistic Effects	989
34.2.6	Antenna Phase Center Calibrations	989
34.2.7	Phase Wind-Up	989
34.2.8	GNSS Transmitter Models and Information	990
34.2.9	Models in Downstream Applications ..	991
34.3	POD Process	992
34.4	Estimation Strategies	993
34.4.1	Estimators	993
34.4.2	Parameterization	994
34.4.3	Ground Stations	995
34.4.4	GNSS Orbits	995
34.4.5	Clock Offsets	996
34.4.6	Earth Orientation	996
34.4.7	Phase Ambiguity Resolution	996
34.4.8	Multi-GNSS Processing	997
34.4.9	Terrestrial Reference Frame	998
34.4.10	Sample Parameterizations	998
34.4.11	Reducing Computation Cost	999
34.5	Software	1000
34.6	Products	1001
34.6.1	IGS Orbit and Clock Combination	1002
34.6.2	Formats and Transmission	1004
34.6.3	Using Products	1005
34.7	Outlook	1005
	References	1006

Most applications of GNSS rely on knowledge of the orbital positions and clock offsets of the transmitter satellites. These parameters are obtained from an estimation process that combines tracking data from terrestrial stations, measurement models, and satellite force models to adjust a set of parameters representing station positions, atmospheric delays, satellite orbits, clock offsets, and the Earth's orientation.

There are broadly three communities performing routine orbit and clock determination of the GNSS

constellations: the control segments, the analysis centers (ACs) of the International GNSS Service (IGS; Chap. 33), and commercial services. The GNSS operational control segments perform orbit and clock determination in real time or near real time using tracking data from a limited set of highly secure ground stations. The orbit and clock solutions are then predicted forward and transmitted to users via the navigation message for real-time use. While the Global Positioning System (GPS) and Global'naya Navigatsionnaya Sput-

nikova Sistema (GLONASS) broadcast orbit and clock have been very good in recent years, they are nevertheless significantly less accurate than solutions produced by the IGS or comparable precise services [34.1]. The main reason for this is that the broadcast orbit and clock is predicted based on a *zero age of data* solution derived from a small tracking network. The prediction time-span can be significant: currently for GPS and GLONASS the update interval can be as long as 24 h. For Galileo, the update rate is reduced to about 100 min, which makes more accurate broadcast ephemerides possible. However, the accuracy of broadcast ephemerides is not likely to become sufficient for applications requiring decimeter or better positioning and time transfer, which in turn drives the need for more accurate and precise orbit determination (POD) and clock products.

The main goals of the IGS are to collect and archive GNSS data from a global network, and to generate precise products for the GNSS constellations. The products of the IGS, in particular the orbits, clocks, and station positions give users direct access to the International Terrestrial Reference Frame (ITRF) with accuracies not previously available with such ease. By taking at least one of the stations of the IGS tracking network, the IGS

orbits, and the IGS station position solution, a user can position a network of stations with millimeter precision in the global reference frame. The high quality and public availability of the IGS products have led to the use of GNSS for many new applications, for example, meteorology and time transfer, and most certainly paved the way for commercial high-accuracy GNSS services operating in the market today. Details of the roles and products of the IGS are given in Chap. 33.

Commercial services typically aim to bridge the gap between the control segment solutions, which are very robust but have lower accuracy, and products available from the IGS, which are very accurate but provided on a *best effort* basis. These services generally cater to precise real-time navigation needs of, for example, the maritime and agricultural industries.

This chapter describes what is required for the generation of highly accurate GNSS orbit and clock products for the most demanding applications. We discuss the tracking network, describe relevant models, key GNSS system parameters, estimation strategies, and software implementations. We summarize post-processed GNSS products available from analysis and combination centers, and look at POD for multiple GNSS constellations.

34.1 Global Tracking Network

GNSS POD requires geodetic GNSS tracking stations providing at least dual-frequency measurements to account for ionospheric path delays. A globally distributed tracking station network is needed to ensure adequate observation strength for all satellites at every epoch in the processing arc. This particularly benefits clock offset determination since these parameters are normally not constrained in the estimation. Due to the unknown satellite and receiver clock offsets, and the ambiguous carrier-phase observations, GNSS phase observations contain only limited information regarding the station-satellite geometry. Basically, the carrier-phase observations only provide information about the change in the station-satellite geometry between observation epochs. This lack of information may be overcome – to a certain extent – by ensuring that the satellites are always observed by multiple ground stations. Complete loss of tracking of a satellite, even for only a few epochs, would mean that all the carrier-phase observations will be reset, and thus new ambiguity parameters will have to be estimated. This not only weakens the solutions with regard to orbital parameters,

but also makes it impossible to solve for the transmitter clock offset during the outage.

An interesting question is how many stations are needed to obtain an acceptable high-accuracy solution. The response is driven by the following considerations:

- **Accuracy:** while additional stations in principle improve the accuracy and robustness of the solution, there is a point of marginal return given the precision of the measurements and redundancies in the tracking geometry (we expect accuracy to improve as a function of \sqrt{n} , where n is the number of observations).
- **Computational expense:** more stations lead to longer computation times, typically increasing by a factor of p^2 , where p is the number of estimated parameters. This is a key consideration for services with stringent latency requirements, as fewer stations allow for faster processing.
- **Station costs:** more stations mean more costs for equipment installation and data transmission. This is an important factor for services relying on pro-

proprietary networks (GNSS operators and commercial providers).

To investigate the impacts of the size of the network, we computed GPS POD solutions for eight consecutive days using real data from a network of 20 to 100 stations in steps of five. The obtained orbit quality was determined by comparing the resulting orbit to the IGS final orbit for the days in question. Figure 34.1 plots the median of the (absolute) residuals of the orbit differences as a function of the number of stations that were used in the solution. When increasing the number of stations, we kept the previous network and added new stations.

The results in Fig. 34.1 indicate that 60 stations should be sufficient to reach IGS orbit quality. A further increase in the number of stations does not significantly improve the orbit quality, at least not as reflected in the median of the squared residuals. So we may conclude that 60 stations are sufficient to achieve the accuracy of

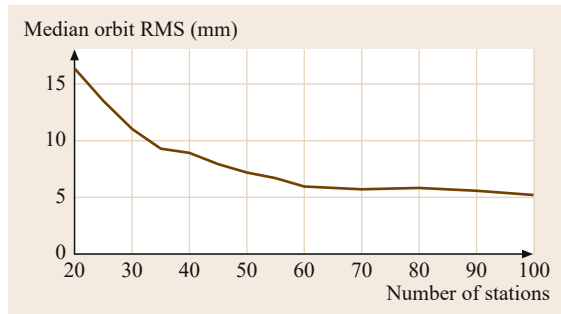


Fig. 34.1 Relation between the number of stations and the satellite orbit quality compared to the final orbits of the IGS

the IGS Rapid or Final products, as long as the tracking network is sufficiently well distributed to track all satellites from at least a handful of stations at any given epoch.

34.2 Models

GNSS software developers are well aware of meter level corrections that must be applied to range observations to eliminate effects such as special and general relativity, clock offsets, and atmospheric delays. All these effects are quite large, exceeding several meters, and must be considered even for pseudorange-only applications. When combining satellite positions and clocks precise to a few centimeters with ionospheric-free carrier-phase observations (few millimeter resolution), it becomes important to apply additional corrections that may not need to be considered in pseudorange or even differential phase processing.

This section describes relevant models given in the IERS conventions [34.2], followed by discussions of antenna phase center calibrations and key parameters related to the GNSS spacecraft. An overview of the important models is given in Table 34.1. References to sections of this handbook with more detailed information are given in the right column.

34.2.1 Reference Frame Transformation

The orbit dynamics of GNSS satellites is usually modeled in an Earth-centered inertial (ECI) frame (Sect. 3.2). However, station and satellite positions are conventionally expressed in an Earth-centered Earth-fixed (ECEF) reference frame, so a transformation between the ECI and the ECEF frame is necessary. This transformation is traditionally composed of three components (Sect. 2.5):

- Precession and nutation
- Polar motion
- UT1 and length of day (LOD).

In global GNSS solutions, precession and nutation are modeled with the IAU2000A R06 model, whereas polar motion and LOD are usually estimated (Sect. 34.4.6). UT1 cannot be determined by GNSS due to correlations with cross-track components of estimated orbital elements [34.3]. As a consequence, UT1 is fixed to values determined by very long baseline interferometry (VLBI) published in the *Bulletin A* [34.4] or the IERS C04 series [34.5]. As polar motion and LOD parameters are usually estimated with a temporal resolutions of 1 day, subdaily variations in these parameters mainly caused by ocean tides have to be considered with a specific model (Table 34.1).

34.2.2 Site Displacement Effects

Terrestrial stations undergo periodic movements (real or apparent) reaching a few decimeters that are not included in linear terrestrial reference frame (TRF) position models. Details are given in Chap. 2. Since most of the periodic station movements are nearly the same over broad areas of the Earth, they nearly cancel in relative positioning over short (< 100 km) baselines. However, to obtain precise station coordinates consistent with ITRF conventions with longer baselines or undifferenced processing, such station movements must

Table 34.1 Common correction models for GNSS data processing. IERS2010 refers to the International Earth Rotation and Reference Systems Service (IERS) conventions (2010). GPT = Global Pressure and Temperature (model); GMF = Global Mapping Function; VMF = Vienna Mapping Function; PCV = phase center variations; PCO = phase center offset; DCB = differential code bias

Model component	Maximum effect	Model	References
Nutation	± 19 as	IAU2000A R06	Sect. 2.5
Subdaily polar motion	± 1 mas	IERS2010	[34.2]
Subdaily length of day	± 0.7 ms	IERS2010	[34.2]
Plate motion	Up to 1 dm/y	IGb08	[34.6]
Solid Earth tides	Up to 40 cm	IERS2010	[34.7], Sects. 2.3.5, 25.2.3
Ocean tidal loading	1–10 cm		[34.8], Sect. 25.2.3
		FES2004	[34.9]
		FES2012	[34.10]
Solid Earth pole tide	Up to 25 mm	IERS2010	Sects. 2.3.5, 25.2.3
Ocean pole tide loading	Up to 2 mm	IERS2010	[34.11]
Atmospheric tidal loading	Up to 1.5 mm	IERS2010	[34.12]
Troposphere (hydrostatic)	≈ 2.3 m ^a		Sects. 6.2.3, 19.3.2, 25.2.1
		GPT/GMF	[34.13, 14]
		GPT2	[34.15]
		VMF1	[34.16]
Ionosphere (1st order)	Up to 30 m ^b	LC ^c	Sects. 6.3.5, 19.3.1, 25.2.1
Ionosphere (higher order)	0–2 cm	IERS2010, IGRF11 ^d	[34.17], Sect. 25.2.1
Relativistic corrections	Up to ± 7 m ^e	IERS2010	[34.18], Sects. 5.4, 19.2
Satellite antenna z-offsets	0.7–2.7 m	igs08.atx	[34.19], Sect. 25.2.2
Satellite antenna PCVs	Up to 12 mm	igs08.atx	[34.19], Sect. 19.5
Receiver antenna PCOs	Up to 16 cm	igs08.atx	[34.19], Sect. 19.5
Receiver antenna PCVs	Up to 3 cm	igs08.atx	[34.19], Sect. 19.5
Phase wind-up	few cm	[34.20]	Sects. 19.4.1, 25.2.2
GPS satellite L1 C/A P(Y) DCBs	Up to 1 m	cc2noncc ^f	Sect. 19.6.1
Attitude	$\pm 180^\circ$ ^g		[34.21], Sect. 3.4
		GPS: [34.22, 23]	
		GLO: [34.24]	
		BDS: [34.25]	
		QZS: [34.26]	
Albedo	1–2 cm ^h	[34.27]	Sect. 3.2.2
Antenna thrust	5 mm ⁱ	[34.28, 29]	[34.27]
Gravity field	3 km ^j	EGM2008	[34.30]

^a In the zenith direction

^b In the zenith direction for GPS L1 frequency

^c LC is the ionosphere-free linear combination of dual-frequency observations (Sect. 20.2.3)

^d International Geomagnetic Reference Field [34.31]

^e Eccentricity correction for satellite clocks (largest effect)

^f Available at <http://acc.igs.org/>

^g Affects phase center location (instantaneous attitude error) and phase wind-up (accumulated attitude error)

^h For GPS

ⁱ For GPS Block IIA

^j Orbit error after two GPS revolutions when neglecting potential terms > 0 (Sect. 3.2)

be considered. This is accomplished by adding the site displacement correction terms to the linear nominal coordinates. The most significant corrections are summarized next.

Solid Earth Tides

The *solid* Earth is deformed due to the gravitational forces of the Sun and Moon. These solar and lunar tides cause periodic vertical and horizontal site displacements, which can reach about 30 cm and 5 cm in the radial and horizontal directions, respectively (Sect. 2.3.5). There is a latitude-dependent permanent displacement and a periodic displacement with predominantly semidiurnal and diurnal periods of changing amplitudes (Fig. 34.2). The periodic part is largely averaged out for static positioning over 24 h. The permanent part, however, remains. Even when averaging over long periods, neglecting this effect in point positioning would result in systematic position errors of up to 12 cm and 5 cm in the radial and horizontal directions, respectively.

The solid Earth tides may be represented by spherical harmonics of degree and order (n, m) characterized by the Love number h_m and the Shida number l_n . The effective values of these numbers weakly depend on station latitude and tidal frequency, which need to be taken into account when a millimeter-level position solution is desired. Note that the estimated ITRF station positions are corrected for the (conventional) permanent part of the solid Earth tides, resulting in a so-called *conventional* coordinate system.

Tidal Ocean Loading

Ocean loading results mainly from the load of ocean tides on the Earth's crust and is dominated by diurnal and semidiurnal periods (Fig 34.3). Displacements due to tidal ocean loading are almost an order of magnitude smaller than those due to solid Earth tides. Tidal ocean loading is also more localized, and by con-

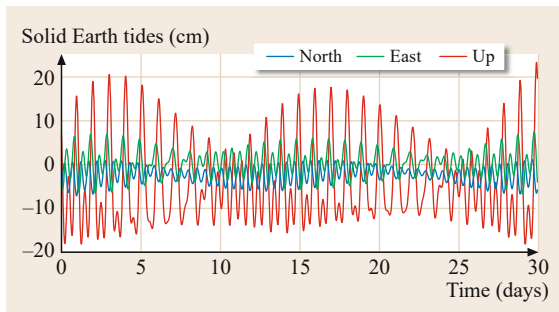


Fig. 34.2 Deformations due to solid Earth tides at Wetzell, Germany

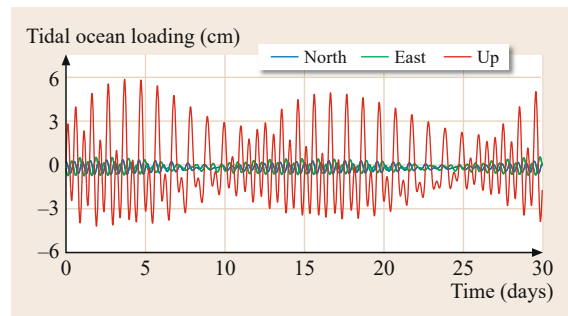


Fig. 34.3 Deformations due to tidal ocean loading at O'Higgins, Antarctica

vention does not have a permanent component. Tidal ocean loading effects can be modeled for any station using the software package provided with the IERS conventions [34.2]. Computation of the station-specific loading effects using HARDISP requires the amplitudes and phases for the radial, south (positive), and west (positive), directions. The amplitudes and phases for any site may be obtained from the on-line ocean loading service of Chalmers University [34.32].

Typically, the M2 amplitudes are the largest and do not exceed 5 cm in the radial and 2 cm in the horizontal directions for coastal stations. For centimeter precision, one should use a recent global ocean tide model, such as FES2004, EOT11a, FES2012, or newer. It may even be necessary to augment the global tidal model with local ocean tides digitized, for example, from local tidal charts. The station-specific amplitudes and phases may also include subdaily center-of-mass (CoM) tidal variations. In that case, ocean loading corrections have to be included for all stations regardless of proximity to an ocean. To be consistent with the subdaily Earth orientation parameter (EOP) convention, the IGS includes subdaily tidal CoM in ocean loading corrections when generating IGS POD solutions.

Pole Tides

Changes in the Earth's spin axis with respect to its crust, that is, the polar motion, cause periodical deformations due to minute changes in the Earth centrifugal potential. The variation of station coordinates caused by the pole tide can amount to around 2 cm and therefore needs to be taken into account. Unlike solid Earth tide and ocean loading effects, the pole tides do not average to nearly zero over 24 h. They are slowly changing according to the polar motion, and predominately vary at seasonal and Chandler (430 days) periods. Polar motion can reach up to 0.8 as and the maximum polar tide displacements can be up to 25 mm in height and 7 mm in the horizontal directions [34.2].

Ocean Pole Tides

The ocean pole tide is generated by the centrifugal effect of polar motion on the oceans. Polar motion is dominated by the 14-month Chandler wobble and annual variations. At these long periods, the ocean pole tide is expected to have an equilibrium response, where the displaced ocean surface is in equilibrium with the forcing equipotential surface. A self-consistent equilibrium model of the ocean pole tide is presented in [34.11]. This model accounts for continental boundaries, mass conservation over the oceans, self-gravitation, and loading of the ocean floor. The load deformation vector is expressed in terms of radial, north, and east components and is a function of the wobble parameters. Given that the amplitude of the wobble parameters is typically of order 0.3 as, the load deformation is typically no larger than about (1.8, 0.5, 0.5) mm in the (radial, north, east) components.

Tidal Atmospheric Loading

The diurnal heating of the atmosphere causes surface pressure oscillations at diurnal S1, semidiurnal S2, and higher harmonics. These atmospheric tides induce periodic motions on the Earth's surface [34.33]. The maximum amplitude of the vertical deformation is 1.5 mm for both the S1 and the S2 component. Being close to the orbital period of the GPS satellites, modeling of the S2 effect is especially important in order to minimize aliasing into dynamic parameters [34.34]. The IERS2010 conventions recommend calculating the station displacement using the S1 and S2 tidal model given by [34.12] (Fig. 34.4). As of 2015, not all IGS analysis centers apply tidal atmospheric loading models but this will likely be harmonized in time for the next IGS reprocessing campaign.

34.2.3 Tropospheric Delay

The nondispersive delay imparted by the atmosphere on a radio signal up to 30 GHz in frequency reaches

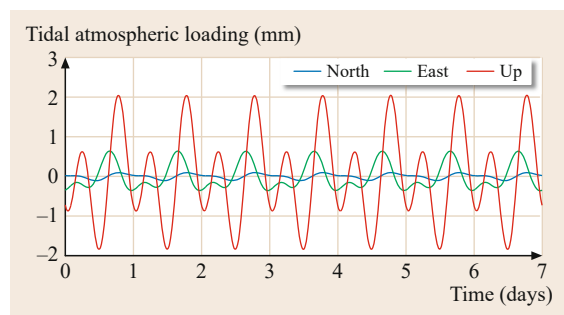


Fig. 34.4 Deformations due to tidal atmospheric loading at Fortaleza, Brazil

a magnitude of about 2.3 m in the zenith direction at sea level [34.35]. It is conveniently divided into hydrostatic and wet components. The hydrostatic delay is caused by the refractivity of the dry gases (mainly N_2 and O_2) in the troposphere and by most of the nondipole component of the water vapor refractivity. The rest of the water vapor refractivity is responsible for most of the wet delay. The hydrostatic delay component accounts for roughly 90% of the total delay at any given site globally, but can vary between 80 and 100% depending on location and time of year. The relation between the delay at zenith direction and the actual observation direction is given by a mapping function (Chap. 6). Common mapping functions are the empirical Global Mapping Function (GMF, [34.14]) and the Vienna Mapping Function 1 (VMF1, [34.36]).

The hydrostatic delay can be accurately computed a priori based on reliable surface pressure data using the formula of [34.37] as given by [34.38]. One source for pressure data are the Global Pressure and Temperature (GPT) model and its successor GPT2 [34.13, 15]. Another possibility is the use of troposphere zenith delays derived from numerical weather models. Global grids with a spatial resolution of $2.0^\circ \times 2.5^\circ$ and a temporal resolution of 6 h obtained from European Centre for Medium-Range Weather Forecasts (ECMWF) data are provided together with the VMF1 coefficients.

There is currently no simple method to derive an accurate a priori value for the wet tropospheric delay, although research continues into the use of external monitoring devices (such as water vapor radiometers) for this purpose. Thus, in precise applications the residual zenith delay is usually estimated. Likewise, horizontal troposphere gradient parameters, needed to account for a systematic component in the North/South direction, are estimated rather than modeled (Sect. 34.4.3).

34.2.4 Ionospheric Delay

GNSS signals are refracted by free electrons and ions in the ionosphere, causing the signals to bend and change speed as they traverse this region. The ionization is caused by rays from the Sun and depends strongly on local time and solar activity. The signal delay due to the ionosphere can vary from meters to tens of meters depending on the ray path and ionospheric activity [34.35]. The delay is dispersive and may be eliminated to first order by linearly combining observations on two or more frequencies (with the side effect of increasing measurement noise, Sect. 20.2.3).

The second-order ionospheric effect can be removed using total electron content (TEC) estimates based on tracking data, an estimated global ionosphere model (GIM), or a climatological model such as the

International Reference Ionosphere [34.39]. The effect can be as large as a few centimeters and should be considered in precise GNSS analyses. An important impact of modeling second-order ionospheric delays is on the recovered terrestrial reference frame, which experiences an apparent southward shift of station coordinates of up to a few millimeters [34.40, 41]. For TRF comparisons spanning several years, this can equate to just over 1 cm in the Helmert transformation z -translation component at the fit epoch [34.42]. Other TRF transformation parameters are negligibly affected. There is also a small impact to satellite orbit positions (a few millimeter shift that is latitude dependent) and up to a 1 cm difference in the transmitter clock estimates [34.43]. Third-order ionospheric delay effects accumulate due to small path differences between signals at different frequencies. These can reach 1 mm in magnitude and are typically ignored in contemporary GNSS processing [34.17, 44]. Another ionospheric effect that is neglected by most IGS ACs is related to ray bending (excess path length) although it can reach a few millimeters at low elevations [34.2].

34.2.5 Relativistic Effects

Relativistic effects relevant for GNSS can be separated into three categories:

- Orbit effects (Sect. 3.2.2)
- Clock effects (Sect. 5.4)
- Propagation effects (Sect. 19.2).

The largest effect is a periodic transmitter clock variation caused by the noncircular orbits of the GNSS satellites. As a result, the satellites' speed and gravitational potential vary with orbital position, introducing a once-per-revolution variation of the transmitter clock with an amplitude of about 23 ns for a GPS satellite with an eccentricity of 0.01 [34.45]. In GNSS POD processing (including IGS products) this effect is modeled by convention, so the published transmitter clocks do not include this term. Smaller effects due to the oblateness of the Earth [34.46] and higher order terms of the Earth's gravity field are usually not considered for the product generation.

34.2.6 Antenna Phase Center Calibrations

Antenna calibrations are critical to high-accuracy GNSS processing. The calibrations define the points in space at which the electromagnetic ranging signal emanates from the transmitter antenna and induces voltage in the receiver antenna. In other words, the measurement geometry refers to the electrical phase centers. These are a function of local azimuth and elevation, fre-

quency, and pseudorange or carrier phase (i.e., group or phase delay). Phase center calibrations are typically separated into a phase center offset (PCO, mean of the total calibration) and a phase center variation (PCV), which varies as a function of azimuth and elevation. In the IGS, absolute calibration standards have been adopted since 2006 [34.47]. The reader is referred to Chap. 17 and Sect. 19.5 for more details.

The transmitter calibrations are needed to refer range measurements to the spacecraft CoM. This is an important link because in POD the modeled spacecraft dynamics and estimated orbits refer to the CoM. The corresponding clock estimates, however, refer to the transmitter antenna phase center. Hence a user of precise products must apply antenna calibrations consistent with those used in the POD solution in order to realize the best accuracy.

Antenna calibration models are closely related to TRF implementation in POD. This is because antenna PCOs in the radial direction for both transmitters and receivers are not separable from TRF scale. IGS standard transmitter antenna calibrations are estimated while keeping scale fixed to a particular ITRF realization. The calibrations are estimated in global POD solutions, with receiver antenna positions fixed to the ITRF and ground calibrations fixed to absolute test range measurements [34.47]. In this manner, the TRF scale is handed off to the transmitter calibrations. It is therefore necessary to apply antenna calibrations consistent with the desired ITRF realization in POD solutions. Furthermore, a consistent set of calibrations must be derived for each version of the ITRF.

34.2.7 Phase Wind-Up

GNSS satellites transmit circularly polarized radio waves, so the observed carrier-phase depends on the mutual orientation of the satellite and receiver antennas. A full rotation of either the receiver or transmitter antenna around its boresight axis will change the carrier-phase measurement by one cycle. This effect is called *phase wind-up* [34.20]. A static receiver antenna remains oriented toward a fixed reference direction (usually north), but the motion of the transmitter relative to its boresight induces wind-up. Furthermore, the transmitter antennas rotate as the satellites yaw about their Earth-pointing axis (coincident with the antenna boresight, see Sect. 34.2.8).

During a satellite eclipse, the rotation can reach up to one revolution within half an hour. Consequently the phase data should be corrected for the wind-up effect. If wind-up is neglected in POD processing, the unmodeled change in carrier phase is absorbed as much as possible in unrelated parameters (Sect. 34.2.8).

34.2.8 GNSS Transmitter Models and Information

High-accuracy POD requires knowledge of and models for many satellite system parameters. These include satellite attitude, physical spacecraft geometry and material properties, operational information regarding the pseudo-random noise (PRN) codes transmitted by each space vehicle, maneuvers, and satellite health.

Spacecraft Attitude

POD requires knowledge of the satellite's attitude to relate range measurements from the antenna phase center to the spacecraft CoM using known antenna calibrations in the spacecraft body system. Nominal GNSS attitude control is guided by two constraints: pointing the transmit antenna toward the Earth's center, and pointing the solar panels, which rotate about their longitudinal axis, toward the Sun in order to obtain maximum energy transfer (Sect. 3.4). As discussed in [34.21], the IGS commonly adopts a right-handed spacecraft coordinate system where the body-fixed z -axis is aligned with the antenna boresight direction, where the y -axis points along the solar panel longitudinal axis, and x completes the right-handed set. Typically, the z -axis is controlled to point to the center of the Earth and the y -axis is kept perpendicular to the Sun direction. Within this so-called yaw-steering mode [34.48], the yaw angle ψ is the angle between the x -axis of the spacecraft and the along-track direction (or, approximately, the velocity vector v). This geometry is illustrated in Fig. 34.5.

To maintain nominal attitude, the spacecraft yaws about its z -axis and rotates the solar panels about y as it traverses the orbit. Because the solar panels can only rotate by 180° , the satellite performs a yaw maneuver at orbit noon and midnight (represented by μ equal to 180° and 0° when the spacecraft is closest and furthest from the Sun, respectively) so the solar panels can track the Sun for the next semi-orbit. The rate of the noon and midnight yaw maneuvers depend on the elevation of the

Sun above the orbital plane of the GNSS satellite, commonly referred to as β (Fig. 34.5). The smaller β , the faster the satellite has to maneuver to maintain nominal attitude.

Due to hardware limitations, GNSS satellites cannot maintain nominal attitude for small β angles. Spacecraft attitude control during these periods differs for each GNSS and even for satellite types within a GNSS. Correct attitude modeling is important because mis-modeled yaw attitude in principle introduces errors in all estimated parameters. Mismodeled attitude in particular affects transmitter clock estimates because yaw maneuvers induce phase wind-up that is essentially common to all stations observing a satellite and therefore induces a clock-like effect. Detailed descriptions of the GPS, GLONASS, Galileo, BeiDou, and Quasi-Zenith Satellite System (QZSS) attitude models are found in Sect. 19.4.2.

Spacecraft Structure

Space vehicle geometry and material properties relate to dynamic force models. After gravity from the Earth, Moon, and Sun, solar radiation pressure (SRP) is the third largest force acting on a satellite orbiting the Earth at GNSS altitudes (Chap. 3). The magnitude and direction of the SRP depends on the satellite attitude and the spacecraft's structural geometry and material properties. Several approaches for dealing with SRP have been devised.

One approach is to solve for empirical accelerations representing SRP and other unmodeled forces in each orbit solution. This strategy originated at the Center for Orbit Determination in Europe (CODE), one of the IGS ACs, and is discussed further in Sect. 34.4.4 below.

A second approach is to generate an empirical SRP model from dynamic fits to precise orbits. The result becomes a background model in POD solution which then estimate only tightly constrained correction factors. The Jet Propulsion Laboratory (JPL) GNSS solar pressure model follows this approach to formulate Fourier expansions representing acceleration due to SRP as a function of the spacecraft type, orbit plane β angle, and satellite orbit angle μ [34.49, 50].

A third approach is to apply ray-tracing techniques to the spacecraft structure and optical material properties. Published models include early work by [34.51], resulting in a set of *ROCK* models for GPS Block I/II, as well as work by [34.52] for the GPS IIR spacecraft. Ray-tracing approaches are attractive because they promise to isolate the effects of SRP from other nonconservative forces, but are burdened by heavy computational loads for their generation and typically limited access to precise physical property information for the GNSS spacecraft.

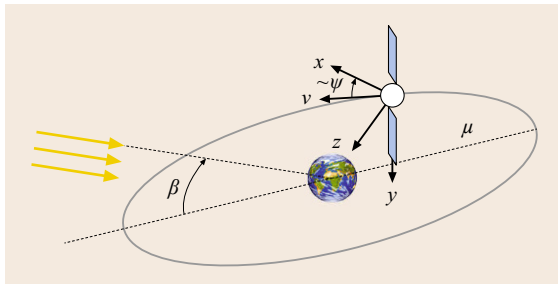


Fig. 34.5 Illustration of spacecraft body-fixed coordinate system and yaw attitude

In addition to SRP, POD nowadays requires modeling of forces on the GNSS spacecraft from optical and infrared Earth radiation, or Earth albedo [34.53]. Albedo radiation explains about half of the few-centimeter height biases seen between radiometric orbits and satellite laser ranging (SLR) measurements [34.54]. Currently all IGS ACs implement albedo visible and infrared models. Most use a model developed by [34.55] that computes albedo forces as a function of spacecraft type, position, time, and Sun position based on measurements of Earth radiation made by the Clouds and the Earth's Radiant Energy System (CERES, [34.56]). The forces due to albedo are largest in the radial orbit component.

Antenna Thrust

GNSS transmitters emit electromagnetic signals with total transmit powers upward of 70 W [34.29]. This results in a *recoil* acceleration in the orbit radial direction, which is now modeled by most IGS ACs, at least for GPS and GLONASS [34.52]. While the impact can be absorbed by an empirical acceleration parameter, modeling this physical effect in principle improves the recovery of transmitter phase centers and clock offsets since these parameters are correlated with antenna thrust.

Operational Information

Operational GNSS information is also needed for POD of GNSS satellites. Receivers typically identify the observations of tracked satellites by the number of their PRN code or the orbital slot number (for GLONASS). However, they do not know the relation between a given PRN/slot and the physical spacecraft, which may vary over time. Complementary knowledge of the unique space vehicle number (SVN) is required for POD because it defines characteristics including attitude control mechanisms, solar pressure and Earth radiation response, clock properties, antenna calibration, signal types, and emitted power. PRN/SVN assignments are tracked by the IGS ACs based on information provided by the GNSS system operators where available (see, e.g., the constellation status websites summarized in Table 34.2).

Knowledge of satellite health status over time, as extracted from the navigation message, is needed be-

Table 34.2 Constellation status and notice advisories for GPS, GLONASS, Galileo, and QZSS. For BeiDou no such official information is available as of early 2016

GNSS	Item	Reference
GPS	Status	[34.57]
	NANU ^a	[34.58]
GLONASS	Status	[34.59]
	NAGU ^b	[34.60]
Galileo	Status	[34.61]
	NAGU ^c	[34.62]
QZSS	Status	[34.63]
	NAQU ^d	[34.64]

^a Notice Advisory to NAVSTAR Users

^b Notice Advisory to GLONASS Users

^c Notice Advisory to Galileo Users

^d Notice Advisory to QZSS Users

cause unhealthy periods may indicate payload maintenance, orbit maneuvers, or nonstandard signal transmission. One typically excludes unhealthy satellites from low-latency, automated processing (e.g., ultra-rapid and rapid) to ensure reliable product delivery, but several IGS ACs include unhealthy satellites in their final products as long as all quality metrics are satisfied. This approach is possible because sometimes satellites are marked unhealthy for reasons unrelated to the navigation signal performance. Such unhealthy periods are usually announced in advance by the GNSS system operators via a so-called Notice Advisory (Table 34.2).

34.2.9 Models in Downstream Applications

It is important to apply consistent models in POD processing and downstream applications utilizing POD products (e.g., precise point positioning (PPP), Chap. 25). In particular, inconsistent antenna calibration and satellite attitude models can result in measurement model errors, which will be absorbed by estimated parameters to the extent possible. The remaining misfit will be evident in postfit residuals. The risk of contaminating physical parameters of interest is, therefore, significant. For this reason, GNSS software providers recommend using POD products generated with the same software as this reduces the likelihood of inconsistent models.

34.3 POD Process

The POD process consists of several discrete steps leading from raw observations through data editing and measurement modeling to parameter estimation and product generation. A high-level diagram showing the connected steps is shown in Fig. 34.6. The top row depicts needed inputs, including raw observation data (typically in Receiver INdependent EXchange (RINEX) format, Annex A.1.2), nominal GNSS orbits/clocks, and EOPs. Satellite and station metadata (PRN/SVN conversions, spacecraft models, nominal station coordinates, antenna information, etc.) are needed at various steps and are not explicitly depicted.

Item 1 in the figure represents the data editing procedure, which evaluates the raw observations and linear data type combinations (e.g., range minus phase, widelane phase, see Chap. 20) to identify poor quality observations (such as very short arcs, inconsistent range/phase) and carrier-phase cycle slips. If the nominal orbits and clocks are of good quality, then data for

each station may be edited using a PPP procedure that includes editing based on raw observations as well as iterative editing based on postfit residuals. The latter approach is generally possible if an ultra-rapid or better nominal orbit and clock product is used as input, and has the advantage of identifying data to remove for each station independently. When postfit residual editing is performed in a global solution, there is some risk that poor quality or nonsensical observations from one station impact postfit residuals across many stations since the global estimation process adjusts parameters to minimize postfit residuals (in a least-squares sense) over all network participants.

Item 2 employs an orbit integrator to generate a dynamic fit to a set of nominal orbits. The nominal orbits could come from the broadcast ephemerides, prior precise solutions, or predictions of past precise orbits to span the processing arc. The dynamic fit is performed iteratively to minimize the difference between the es-

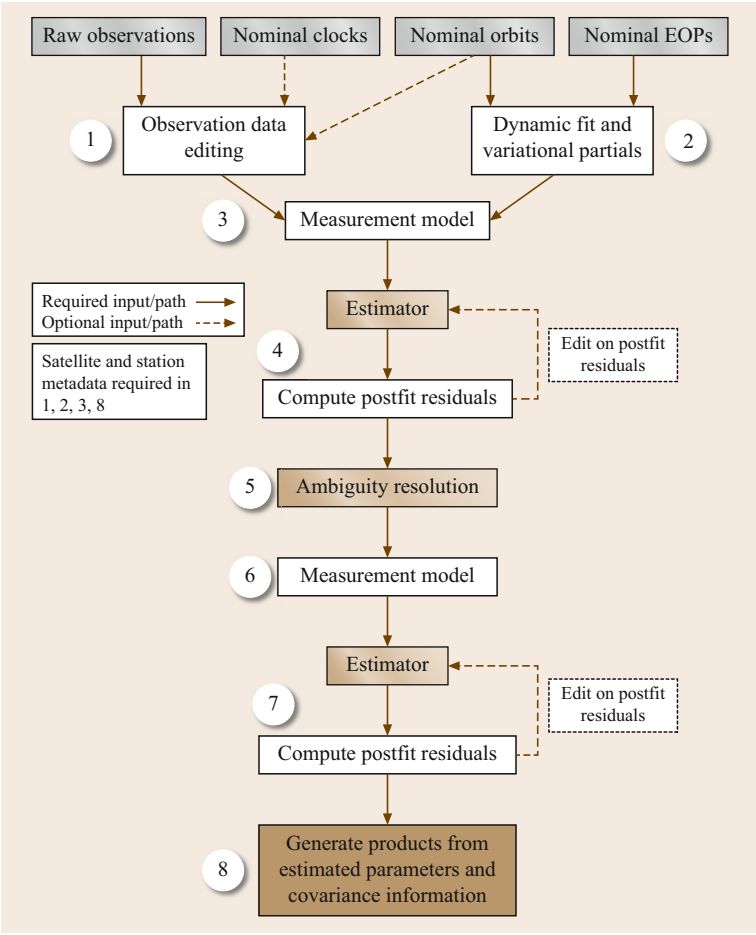


Fig. 34.6 High-level overview of the POD process (representative example)

timated and the nominal orbits using a small set of parameters (at least the epoch state vector or orbital elements as well as a limited number of SRP parameters). The difference between the nominal and generated dynamic orbits reflects both the accuracy of the nominal orbits and the number of fitted parameters. Significant misfits can be used to screen for model errors and satellites not following a sufficiently dynamic trajectory due to, for example, orbit maneuvers or unusual attitude configurations. This step also computes the partial derivatives of the position and velocity vectors with respect to the orbit parameters (variational partials).

The measurement model, labeled as item 3, computes the expected measurements corresponding to the input set of observations (item 1). These are based on instantaneous nominal receiver and transmitter antenna phase centers determined from nominal values and the models described in Sect. 34.2. Since the GNSS measurements (observations) have nonlinear relationships to the estimated parameters (state variables), the observation-state equations are linearized about the nominal state. The computed measurements are subtracted from the observations at this step, and the partial derivatives of the linearized measurement model with respect to the estimated parameters are determined.

The variational partials and outputs of the measurement model are then passed to an estimator that adjusts parameters to minimize a cost function such as the sum of the square of the postfit residuals (item 4). The estimator solves for an adjustment to the nominal state that best fits the observations, typically iterating the solu-

tion until the adjustment is smaller than some threshold (i.e., convergence). This works as long as the nominal state and the estimated state fall within the linear regime of the observation-state relationships. The estimator solution is the sum of the nominal and adjusted parameter values.

The set of estimated parameters usually includes the following: station coordinates, station troposphere delays, transmitter satellite orbits (epoch state, empirical accelerations and/or solar pressure model scales), EOPs, receiver and transmitter clocks, floating point, and integer fixed carrier-phase ambiguities. When more than one GNSS is used, receiver intersystem biases and/or interfrequency biases (IFBs) are additionally estimated. The outputs of the estimator include the parameter adjustments, estimated covariance, and postfit residuals. The residuals may be examined for outliers, and the solution iterated using successively smaller residual outlier thresholds.

The results of this step also provide the input for the subsequent carrier-phase integer ambiguity resolution (item 5, see Sect. 34.4.7 and Chap. 23). After ambiguity resolution, the measurement model is run again to account for resolved integers in the observations (item 6), after which the ambiguity resolved estimator solution is generated (item 7). Here, too, one can iteratively edit observations based on postfit residuals, although almost all outliers should have been removed in prior iterations. The estimated parameters and covariance information is then used to create product files (item 8) for distribution to the users.

34.4 Estimation Strategies

Various estimation strategies are used within the IGS POD community. Some analysis centers process undifferenced pseudorange and carrier-phase observations and estimate all parameters in one integrated solution. Others process double-difference carrier-phase observations, which remove clock parameters from the measurements, and solve for orbits, EOPs, station positions, tropospheric delays, and carrier-phase integer ambiguities. The estimates can then be held fixed in a follow-on solution that uses undifferenced observations to solve for clock parameters consistent with the double-difference solution. The processing arcs found within the IGS range from 24 h to 3 days. Longer arcs are preferred for orbit parameters since additional revolutions improve knowledge of the dynamics. A 1-day arc is preferred for EOP and station position parameters if a daily terrestrial reference frame is of interest. Carrier-phase observations are usually weighted at least

100 times higher than pseudorange due to their significantly higher precision (e.g., 1 cm and 1 m, respectively). Sometimes data from individual stations are also weighted according to the overall root mean square (RMS) level of their postfit residuals in an initial solution, since relatively high postfit residuals may indicate issues with model fidelity or data editing at particular locations. In the following subsections, we describe common estimators, parameterizations, and strategies for additional aspects of POD in further detail.

34.4.1 Estimators

In the IGS community, two primary types of estimators are utilized for GNSS orbit and clock determination. The first is the batch least-squares estimator, which takes all observations, partial derivatives, and a priori covariance for the processing arc to form

a normal equation system that is inverted to compute state variable adjustments and covariance information (state variable uncertainties and correlations). The second category comprises sequential estimators (such as the Kalman or square root information filter [34.65]), which ingest observations one epoch at a time to produce the best state adjustment based on measurements processed so far. Both types of estimators are discussed in detail in Chap. 22. We also refer the reader to [34.66, 67] for thorough reviews of filtering techniques in the context of range measurement processing.

A significant difference between the batch least-squares estimation and Kalman filtering lies in the treatment of stochastic, time-variable parameters [34.67]. In the case of the batch estimator, distinct estimation parameters need to be set up for each new epoch or time interval (e.g., a clock offset at each epoch, a zenith troposphere delay every hour, and daily station coordinates). This creates large normal equation matrices (Sect. 34.4.10), so parameter elimination and back-substitution techniques are used to reduce the computational burden [34.68]. For the Kalman filter, the number of parameters remains constant, but process noise is applied between epochs. The process noise can be configured to apply white noise (no correlation from estimate to estimate), colored noise (correlation over a period of time), or random walk (infinite correlation) updates to a parameter.

34.4.2 Parameterization

All state-of-the-art precise GNSS software packages make available a variety of parameterizations. The most common are as follows:

- *Offset*: adjusts a single, constant value over the processing arc. Commonly used for station coordinates since geodetic stations are considered static over processing arcs of up to several days. Also used for differential code biases (DCBs) (spanning up to monthly intervals) and antenna PCOs.
- *Piecewise-constant*: offset parameters with discrete steps. Typically used for carrier-phase ambiguities.
- *Piecewise-linear*: described by an offset and a slope. Typically used for EOPs.
- *Continuous piecewise-linear*: a piecewise-linear parameterization with imposed continuity at the interval boundaries. It is typically achieved by estimating the parameter values at discrete nodal points located at these boundaries. Compared to the piecewise-linear representation, the number of estimation parameters is reduced by $n - 1$ with n being the number of time intervals. Equivalent to a piecewise-linear representation with a tight continuity constraint. No discontinuities occur inside the processing interval allowing for a more physical parameter representation. The piecewise-linear representation can be transformed to a continuous piecewise-linear representation but not vice versa. Used for station troposphere delays, EOPs in high-rate or multiday solutions, and nadir-dependent satellite antenna PCVs.
- *Epoch independent*: an independent parameter values is estimated at each epoch, for example, used for receiver and transmitter clocks. Equivalent to a constant parameter with a stochastic white noise reset applied at every epoch.

An illustration of these five types of parameterizations is shown in Fig. 34.7.

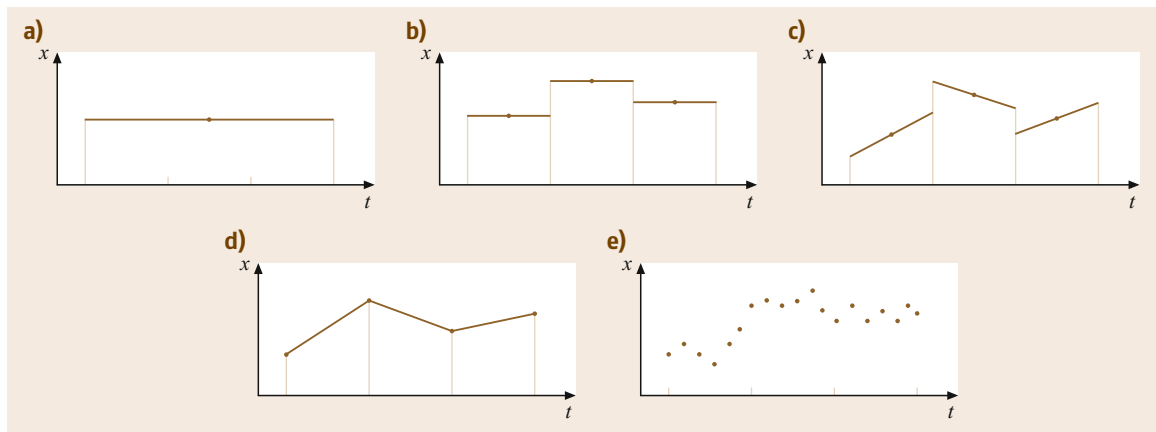


Fig. 34.7a–e Parameterizations used in GNSS data processing: (a) offset, (b) piecewise-constant, (c) piecewise-linear, (d) continuous piecewise-linear, (e) epoch independent

34.4.3 Ground Stations

For ground stations one must estimate coordinates, tropospheric delay, and receiver clock parameters. For geodetic stations, it is usually sufficient to estimate a constant position over the processing arc with loose a priori constraints in the range of meters to kilometers.

The troposphere is modeled using a zenith path delay, a mapping function which may vary by time and the station location (e.g., Global Mapping Function, [34.14]), and horizontal gradient parameters. Although the analysis of observations at low elevations is degraded by increased noise, multipath, and model deficiencies, these observations are important for a decorrelation of troposphere zenith delays, receiver clock, and station height ([34.69] and Fig. 6.3). Therefore, low-elevation data are included but downweighted. Common weighting functions are $1/\sin e$ and $1/\sin^2 e$ with the satellite elevation e , additional weighting functions are discussed in [34.70] and [34.71]. Nevertheless, data at very low elevations are excluded by applying an elevation mask. Typical elevation cut-off angles range from 3° to 10° .

Horizontal tropospheric gradient parameters are needed to account for a systematic component in the north/south direction toward the equator due to the atmospheric bulge [34.72] with magnitudes of about -0.5 and $+0.5$ mm at mid-latitudes in the Northern and Southern hemispheres, respectively. The gradients are generally parameterized for each station as components of a sinusoid varying in azimuth. The gradients also capture the effects of random components in both directions due to weather systems. Failing to model gradients in radiometric analyses can lead to systematic errors in the scale of the estimated terrestrial reference frame at the level of about 1 ppb, and cause latitude and declination offsets in station and transmitter positions [34.73].

The zenith path delay parameter may be estimated as a piecewise constant (random walk) process with an a priori sigma of tens of centimeters and process noise < 1 mm over 5 min (1σ). Likewise gradient parameters, usually the in-phase and quadrature components of an empirical sinusoid fit, may be modeled as random walk processes with similar a priori sigmas and process noise an order of magnitude smaller than the zenith delay. The estimation of receiver clock parameters is discussed in Sect. 34.4.5.

34.4.4 GNSS Orbits

The satellite orbits are estimated using a *reduced dynamic* approach [34.74]. The basic outline is to solve for an epoch state vector (position and velocity or set of osculating elements at a reference epoch), empirical ac-

celerations and/or model scale factors to absorb force model errors (mainly solar radiation, Earth radiation, spacecraft thermal radiation, and transmitter antenna thrust), and, optionally, yaw attitude parameters. For SRP modeling, the strategies currently represented in the IGS analysis community may be broadly divided into two categories.

The first is the *CODE approach* described in [34.75, 76]. Here, one estimates an epoch state plus constant and per-revolution accelerations in the DYB-frame (where D refers to the spacecraft-Sun direction, Y to the body-fixed solar panel axis, and B completes the right-handed set). The classic ECOM (Empirical CODE Orbit Model) approach solves for 6 epoch state parameters and 5 empirical accelerations as constant terms in D, Y, and B and 1/rev terms in B (Sect. 3.2.4). In 2015, the CODE AC updated this strategy to also include two- and four-times per revolution accelerations in the spacecraft-Sun direction (ECOM-2), as this was empirically found to reduce signals related to the GNSS draconitic period seen in terrestrial reference frame transformation parameters [34.77].

The second category, developed at JPL, emphasizes high-fidelity a priori SRP models and estimates tightly constrained SRP model scale and empirical accelerations for each arc [34.49, 50]. A typical set of parameters is the epoch state, an overall constant solar pressure model scale factor, tightly constrained stochastic solar scale in the spacecraft body-fixed x - and z -components, plus constant and constrained stochastic accelerations in y (accounting for unmodeled thermal radiation and SRP forces).

Contemporary orbit solutions submitted to the IGS are derived using both approaches and variations thereof. They show agreement within the expected precision of the estimates. In other words, the approaches independently validate one another and produce high-quality solutions. There are, of course, some advantages and disadvantages for each. The CODE approach has the advantage that it requires no a priori model for SRP forces as they are simply absorbed by the empirical accelerations. It is therefore well suited to dealing with new spacecraft types as soon as tracking data are available. The disadvantage of the approach is that significant empirical forces must be estimated, and there is the risk of mixing dynamical parameters with other physical parameters such as EOPs and geocenter [34.77, 78].

The advantage of the second category is that the forces acting on the satellite are defined by a high-fidelity a priori model such that empirical accelerations can be tightly constrained to account for hopefully small deficiencies in the background models. This, in principle, allows for less mixing of spacecraft dynamics into other geodetic parameters. The disadvantage of

this approach is that it relies on a prior set of precise orbits (from which the SRP model is generated) so it cannot be readily applied to new satellites. Further, any systematic errors in the prior orbit set can affect the resulting SRP model.

The estimation of yaw rates for orbit shadow and noon maneuvers for the GPS Block II/IIA/IIF satellites is also beneficial. In the case of the Block II/IIA spacecraft, this is needed because the maneuver is controlled by analog sensors resulting in yaw rates that can differ from nominal values by as much as 25% [34.48]. While the Block IIF attitude is deterministic, nonetheless discrepancies between nominal and actual attitude for some β angle regimes have been found [34.24]. A new yaw parameter should be setup (or stochastic update performed) for each yaw maneuver since they are independent events. Analyses have shown that GPS Block IIR and GLONASS-M attitude may be accurately modeled without the need for empirical parameters [34.22, 24].

34.4.5 Clock Offsets

The next class of parameters are receiver and transmitter clock offsets. The clocks are generally modeled as unconstrained epoch-independent parameters (no correlation from one epoch to the next). This parameterization makes the solution independent of the quality of the individual station clocks, most of which are not tied to an atomic reference.

Transmitter clocks are driven by atomic frequency standards, but may exhibit discontinuities that make modeling them as, for example, a quadratic function unreliable. The solution is singular if all clocks are left unconstrained, so a reference must be selected. In practice this implies that one holds a particular clock offset or an ensemble of clocks fixed, meaning all other clock offsets are estimated relative to the reference. Generally one can use a station stably tied to a reference time scale such as the coordinated universal time (UTC) realizations of national timing labs.

An alternate approach less sensitive to data gaps is to apply an overall zero mean constraint to an ensemble of clock offsets. Some ACs, for instance, apply a zero-mean constraint to the transmitter clocks. The resulting orbit and clock solution is internally consistent and yields valid formal error estimates for all parameters. The timescale of the solution may be adjusted after the fact by removing a single reference clock offset at each epoch from all clock estimates based on a prioritized list of receiver clocks aligned to GPS time.

The GPS clock parameters provided by the IGS refer (by convention) to the ionosphere-free linear combination of the L1 P(Y) and L2 P(Y) signals. However,

several geodetic GPS receivers only track the C/A code on L1. Therefore, DCBs between the different types of signals have to be considered (Sect. 19.6.1). One can either estimate the DCB parameters, as done by the CODE AC, or one can correct for the DCBs, for example, with the `cc2noncc` tool utilizing the CODE DCB estimates [34.79]. Users of IGS clock products must ensure consistency between the clock parameters on the one hand and the observation types on the other hand by applying the corresponding DCBs. Further sources for DCBs are the Time Group Delay (TGD) parameters of the GPS navigation message (Sect. 7.4.3), the Inter Signal Corrections (ISCs) of the GPS Civil Navigation Message CNAV [34.80], and the IGS MGEX DCB product [34.81].

34.4.6 Earth Orientation

Earth orientation parameters relate the terrestrial reference frame to the celestial reference frame. UT1–UTC measures the rotation rate of the Earth relative to an atomic time scale. The period of one Earth revolution is not constant in time, and rotation time in excess of 24 h is referred to as the length of day (LOD). Global GNSS solutions are sensitive to changes in the Earth rotation rate over a processing arc but not the absolute rotational alignment at the start of the arc. From the change in rotation one can compute LOD as described in [34.2]. The x - and y -coordinates of the Earth rotation axis, as well as their rates, can also be determined from GNSS. By convention the estimated pole coordinates are with respect to the IERS Reference Pole [34.2].

34.4.7 Phase Ambiguity Resolution

To achieve the best precision and accuracy, one should resolve phase measurement integer cycle ambiguities. Details on ambiguity resolution algorithms are provided in Chap. 23, here we focus on the steps taken in a global POD solution.

The process usually starts with a solution that estimates the float-valued ambiguities of the ionosphere-free combination of dual-frequency carrier-phase observations along with the station, troposphere, satellite orbit, clock, Earth orientation, and DCB parameters. For a daily solution with 60 or more well-distributed ground stations one can expect orbit accuracies around 10 cm (3D RMS) and clock accuracies of about 10 cm.

This accuracy is generally sufficient to facilitate ambiguity resolution in a second step, where the above parameters are introduced as known parameters (with given accuracy) to setup systems of equations for resolving double-difference ambiguities for individual pairs of stations. The specific techniques employed in

this step differ widely between analysis centers and software packages [34.82–84], although many of them are based on a widelane/narrowlane approach utilizing the Melbourne–Wübbena linear combination.

Given the wealth of observations available, it is usually best to apply conservative thresholds for accepting ambiguities. It is far better to resolve fewer ambiguities correctly than resolving more ambiguities and resolving an integer incorrectly. This is critical since the integers are introduced as fixed values in a follow-up estimation that effectively treats carrier-phase measurements as highly precise, unbiased ranges. Typically one can expect to resolve upward of 90% of ambiguities in a 60+ station global solution.

34.4.8 Multi-GNSS Processing

Although the orbital configurations of GLONASS, Galileo, and BeiDou satellites in medium Earth orbit (MEO) are similar to GPS, the existing estimation strategies partly fail for these satellites.

Systematic errors have been revealed at the 20 cm level in the Galileo orbit and clock products of four different ACs [34.85]. The stretched shape of the Galileo satellites have been identified as the root cause of these errors and an a priori box model has been developed, significantly reducing the systematic errors [34.86]. A similar orbit quality can also be achieved with the newly developed ECOM-2 model [34.77]. While spacecraft attitude and antenna phase centers differ for these satellites they can in principle be derived using procedures developed for GPS. POD of geostationary satellites as employed by BeiDou is still challenging due to the small changes in observation geometry and frequent maneuvers [34.87], although some progress has been made with alternative orbit parameterizations [34.88].

The estimation of clocks for these systems is also similar to GPS in that the offsets can be treated as unconstrained epoch-independent parameters. In a solution involving GLONASS several additional parameters are, however, needed. First, one must estimate a GLONASS intersystem bias at each receiver. GNSS receiver data typically refer the measurements to the respective system time, while the POD solution refers all estimates to GPS time (by convention). So an overall clock bias captures the difference between the GPS and GLONASS timescales at each receiver. It affects all GLONASS range and phase measurements equally. The estimated values should, in general, be in the vicinity of the GPS–GLONASS system time offset, which today is better than 1 μ s after accounting for the constant 3 h UTC(USNO)–UTC(SU) bias [34.89].

A second set of additional parameters needed for GLONASS are IFBs along each receiver–transmitter

pseudorange link. These link biases are necessitated by the frequency division multiple access (FDMA) architecture of the GLONASS system (Sect. 8.2.2), as the hardware delays experienced by GLONASS signals traveling through the receiving equipment are dispersive. The delays vary due to both the physical equipment as well as environmental factors such as temperature. A reasonable strategy is to estimate each IFB as a constant parameter over each day. Due to the estimation of IFB parameters, the choice of the reference signals for GLONASS clock estimation (e.g., P- or C/A-code) is arbitrary.

The magnitude of the IFB parameters is in the range of decimeters to 3 m depending on the receiver type [34.90]. One must choose a reference for both the system time offsets and IFBs or the solution is singular. One can fix to calibrated values or artificially set these biases to zero for a particular receiver. Either way the solution yields a consistent set of GPS and GLONASS clock offsets suitable for use in combined GPS/GLONASS point positioning; however one must take care to also estimate a GLONASS time offset and IFBs in that solution.

It is clear that the GLONASS clock estimates depend strongly on the choice of the system time offset and IFB reference, as well as the choice of receiver network in general since the IFBs essentially bias the pseudorange on each link (unlike GPS where IFBs are not needed since all signals are on common frequencies). For these reasons, it is quite complicated to compare GLONASS clock estimates from different solutions if a different reference and receiver networks were used. This is the main reason the IGS does not currently produce a GLONASS clock combination [34.91].

Galileo and BeiDou are code division multiple access (CDMA) systems like GPS. Thus, in a combined solution one only needs to estimate GPS to Galileo and/or BeiDou time offsets at each station, parameterized as an overall constellation bias (intersystem bias) over the processing arc. A key choice related to the bias magnitudes is the ionosphere-free observable type. For Galileo an E1/E5a convention is emerging [34.85], while for the BeiDou B1/B2 signals are used [34.92].

An important consideration in multi-GNSS solutions is the weighting of observations from each GNSS. Weighting observations in the same manner regardless of GNSS is of course one possibility. Two other commonly used approaches are to downweight observations from non-GPS constellations, or to derive station clock and troposphere parameters with GPS alone and hold the results fixed in a follow-on solution that estimates intersystem biases, transmitter clocks, IFBs, and satellite orbits for one or more additional constellations. This is beneficial since the satellite constellations and

ground networks for the other GNSSs are (currently) not as large or well distributed as for GPS. Also, detailed models for the newer constellations are still being developed and refined. This situation will change rapidly in the next few years and it is likely that new weighting strategies will emerge.

34.4.9 Terrestrial Reference Frame

A stable, accurate, and well-maintained global TRF is a prerequisite for precise orbit and clock determination and its applications. The TRF underpins POD by defining the origin from which receiver and transmitter locations are defined. It furthermore establishes the framework upon which geophysical processes, such as solid Earth tides or vertical motion due to ocean loading, are modeled and analyzed. Details on the definition and realization of terrestrial reference systems are given in Sect. 2.3.

The most recent version of the International Terrestrial Reference Frame is ITRF2008. However, ITRF2008 is usually not directly used in GNSS applications as the predecessor of the current antenna model was applied for the IGS contribution to ITRF2008. Therefore, a GNSS-only TRF called IGS08 was computed [34.6]. It is aligned to ITRF2008 but includes station-specific corrections to account for the antenna calibration differences. In 2012, an updated version of IGS08 called IGB08 [34.93] was released as the coordinates of more than 30 reference frame stations were degraded due to station displacements induced by earthquakes or equipment changes.

The TRF is realized in global POD solutions in one of three ways:

1. Fixing or tightly constraining a set of station positions to the values defined by the TRF. In principle, three fixed stations are sufficient, although in practice IGS ACs hold at least two dozen globally distributed stations fixed. The positions of the remaining stations in the solution, as well as the GNSS orbits, will therefore be estimated relative to the realization of the TRF defined by the subset of fixed stations. However, fixing or constraining more stations than necessary might result in distortions of the network geometry. This approach is typically used only for low-latency solutions (i. e., ultra-rapid) where realizing a TRF a posteriori is time consuming or not valuable for user applications.
2. Applying *minimum constraints* for a selected set of (core) stations w.r.t. the a priori TRF. For global GNSS solutions, a no-net-rotation condition is mandatory. If the origin of the tracking network (geocenter) is estimated, an additional no-

Table 34.3 Number of observations and estimation parameters in a global GPS solution with 32 satellites and 160 stations. It is assumed that 10 satellites are visible per station and observation epoch

	Sampling	No. of obs./par.
Observations	5 min	460 800
Station coordinates	24 h	480
Troposphere zenith delays	2 h	2080
Troposphere gradients	24 h	640
Orbit parameters	24 h	576
Earth orientation parameters	24 h	5
Ambiguities	Dep. on data	≈ 10 000
Satellite clocks	5 min	9216
Receiver clocks	5 min	46 080
Total number of parameters		≈ 69 000

net-translation condition has to be applied. The advantage of this approach is that the inner geometry of the network is not distorted.

3. Estimation of all station positions in the global solutions with loose a priori constraints. Other than relating to the TRF through antenna calibrations (discussed below), this type of *fiducial free* POD solution does not provide orbits and clocks in a particular TRF but instead realizes a unique frame for that solution. This solution’s frame is likely to exhibit notable rotations with respect to ITRF because the GNSS technique is insensitive to rotating the entire observation geometry. One can compute a best fit Helmert transformation (Chap. 2) for the estimated ground network relative to ITRF using the set of overlapping stations (XYZ translations, XYZ rotations, and scale) and apply this transformation to the orbit solution to place it in the ITRF frame.

The products resulting from one of these approaches are provided in the underlying TRF and therefore directly transfer the TRF to users.

Due to orbit dynamics, the GNSS satellites orbits refer to the CoM of the total Earth system including the oceans and the atmosphere. Tides cause periodic variations in the CoM of the oceans and the atmosphere. This so-called geocenter motion can reach up to 1 cm for diurnal and semidiurnal ocean tides [34.94]. If a corresponding CoM correction (CMC) is applied, the orbits refer to a crust-based Center-of-Network (CoN) frame, otherwise to a CoM frame. Within the IGS, the CoN frame is used for the orbits as well as the clocks [34.95].

34.4.10 Sample Parameterizations

Table 34.3 gives an overview of the estimated parameters and their sampling for a global GPS solution with

Table 34.4 Sample receiver station parameterization (JPL approach) (after [34.50])

Parameter	Configuration	σ_{apr}	σ process noise
Station coordinates (all or a subset of stations)	Offset	1 km	–
Station zenith wet troposphere	Random walk, 10 min updates	0.5 m	$0.03 \text{ mm s}^{-1/2}$
Station gradient wet troposphere	Random walk, 10 min updates	0.5 m	$0.003 \text{ mm s}^{-1/2}$
Station clock offset	White noise, update each epoch	1 s	1 s

Table 34.5 Sample satellite parameterization (CODE approach) (after [34.76, 96])

Parameter	Configuration	σ_{apr}
Keplerian elements at epoch	Offset	No constraint
Acceleration in the Sun-direction (D)	Offset	No constraint
Acceleration in the Y-direction	Offset	No constraint
Acceleration in the B-direction	Offset	No constraint
Acceleration in the B-direction (once per revolution)	Offset	No constraint
Constant radial velocity change	Every 12 h	$1 \cdot 10^{-6} \text{ m s}^{-1}$
Constant along-track velocity change	Every 12 h	$1 \cdot 10^{-5} \text{ m s}^{-1}$
Constant cross-track velocity change	Every 12 h	$1 \cdot 10^{-8} \text{ m s}^{-1}$

Table 34.6 Sample satellite parameterization (JPL approach) (after [34.50])

Parameter	Configuration	σ_{apr}	σ process noise
Position at epoch	Offset	1 km	–
Velocity at epoch	Offset	1 cm s^{-1}	–
Y acceleration	Offset	1 nm s^{-2}	–
Y acceleration	Colored noise, 4 h correlation, updated every 1 h	0.01 nm s^{-2}	$0.0002 \text{ nm s}^{-2} \text{ s}^{-1/2}$
SRP model scale	Offset	1.0	–
SRP model scale in X and Z	Colored noise, 4 h correlation, updated every 1 h	0.01	$0.0002 \text{ s}^{-1/2}$
GPS Block II/IIA yaw rates	Offset per eclipsing midnight/noon turn	0.01 deg s^{-1}	–
Transmitter clock offset	White noise, update each epoch	1 s	1 s

a full constellation of 32 satellites and a network of 160 terrestrial stations at a 5 min sampling rate. Clock offsets make up the vast majority of parameters. These parameters are removed if double-difference measurements are processed. The next largest parameter group is for ambiguities whose number strongly depends on the data quality. Efficient methods to deal with the huge number of almost 70 000 estimation parameters are discussed in Sect. 34.4.11.

More specific sets of estimated parameters are discussed in the following. We list sample a priori standard deviations (σ) and stochastic properties, which provide reasonable solutions based on POD strategies described in [34.50, 75, 76]. We note that some software packages apply stochastic updates directly in a Kalman filter [34.67], while those using a batch filter typically configure a new (optionally constrained) parameter for each update and apply elimination techniques to the normal equations to reduce the computational burden. For the purposes of this discussion we consider both approaches to yield equivalent parameterizations. Table 34.4 shows a sample receiver station parameter

Table 34.7 Sample Earth orientation parameterization (JPL approach) (after [34.50])

Parameter	Configuration	σ_{apr}
X and Y pole	Offset	5 m (relative to a priori)
UT1-UTC rate per arc	Offset	$3.5 \cdot 10^{-8} \text{ s/s}$

configuration, Tables 34.5 and 34.6 show possible satellite parameterizations, and Table 34.7 refers to EOPs. Range and phase data are often given a priori measurement sigmas of 1 m and 1 cm, respectively, or a factor of 1:100, to give credit to the high precision of the carrier-phase measurements. This in effect produces a phase-based solution that aligns the clock estimates to the pseudoranges.

34.4.11 Reducing Computation Cost

A given arc and dataset can of course be processed as one solution from start to finish. This approach is often taken to produce few-hour latency (ultra-rapid) or next-day (rapid) solutions. There are, however, trade-

offs between accuracy, processing time, and the number of stations and orbiters processed, as discussed in Sect. 34.1. A few strategies are commonly used to maximize the number of receivers and transmitters while minimizing processing time:

- Real-time clock estimation (few second latency) based on ultra-rapid orbits: GNSS orbits can be predicted from a precise solution with sufficient accuracy (over many hours and even days) that they may be input and held fixed in a real-time GNSS clock filter. This reduces the computational burden and complexity of the real-time process, which estimates only clock and troposphere parameters. This approach is widely used to produce real-time GNSS solutions [34.97].
- Reuse of normal equation systems from prior POD solutions: many software packages include tools to manipulate and stack normal equation systems, and take advantage of these capabilities to minimize new computations for low-latency processing. For instance, utilizing prior rapid or ultra-rapid normal equation systems and appending a few hours of new data to generate an ultra-rapid solution. This avoids recomputing the measurement model and partial derivatives for a significant portion of the dataset. Normal equation stacking is also an efficient tool to generate multiday orbital arcs.
- Stacking normal equation systems for a set of non-overlapping networks: this approach is sometimes

used where it is desired to process as many stations as possible (e.g., to contribute to the TRF). This is a parallelization technique as the subnetwork normal equations are generated on separate systems prior to stacking and inversion.

- Double-difference processing: some software packages process double-difference observable combinations to eliminate transmitter and receiver clock offsets from the observations. This reduces the number of parameters at each epoch significantly but still provides access to orbit, EOP, atmospheric delay, and station position parameters. Many science applications of GNSS have no interest in clock offsets and benefit from a reduced computational burden when processing double difference observables.
- Clock densification: high-rate satellite clock parameters required for highly dynamic applications (e.g., kinematic ground stations or kinematic orbit determination of low Earth orbiters, Chap. 32) dramatically increase the number of unknown parameters. For the example network of Table 34.3, more than 500 000 clock parameters would have to be estimated for 30 s sampling. In order to save computation time [34.98] developed a method utilizing epoch-differenced phase observations. This *efficient high-rate clock interpolation (EHRI)* algorithm densifies the 5 min satellite clock parameters obtained from the global clock estimation to 30 s or even 5 s sampling and reduces computation time by a factor of about 10.

34.5 Software

A variety of software packages for GNSS POD have been developed at academic, research, and commercial institutions. We give brief descriptions of some of these below and summarize the products currently produced by each:

- Bernese GNSS Software, developed at the Astronomical Institute of the University of Bern ([AIUB](#)). An extensive software package, Bernese is used for GNSS and low Earth orbit ([LEO](#)) POD, precise point positioning, estimation of DCBs, antenna calibrations, ionosphere and troposphere estimation, and more. AIUB along with other institutions make up the Center for Orbit Determination in Europe (CODE), which contribute GPS and GLONASS products (orbits, clocks, DCBs, antenna calibrations, troposphere and ionosphere solutions) to the IGS, as well as multi-GNSS products including Galileo, BeiDou, and QZSS to the IGS MGEX [34.68].
- NAPEOS (Navigation Package for Earth Orbiting Satellites), developed at the European Space Agency ([ESA](#)). NAPEOS is used for GNSS and LEO POD, precise point positioning, estimation of DCBs, antenna calibrations, ionosphere and troposphere parameters, etc. NAPEOS is used to generate products at the ESA IGS AC, which contributes GPS and GLONASS POD as well as troposphere and ionosphere products to the IGS [34.99].
- GIPSY (GNSS-Inferred Positioning System and Orbit Analysis Simulation Software) is developed at the National Aeronautics and Space Administration ([NASA](#)) Jet Propulsion Laboratory ([JPL](#)). GIPSY is used to generate GPS, GLONASS, and LEO orbit and clock products, PPP solutions, ionosphere and troposphere parameters, and produces the JPL AC contributions to the IGS. GIPSY processing supports NASA flight missions (LEOs and aircraft positioning) as well as atmospheric calibrations for the Deep Space Network [34.100], among others.

- EPOS-8 (Earth Parameter and Orbit System), developed at Deutsches GeoForschungsZentrum (GFZ), is another package with broad capabilities for GNSS POD, troposphere/ionosphere estimation, transmitter antenna calibrations, etc. GFZ is an IGS AC providing GPS and GLONASS products, as well as multi-GNSS solutions for Galileo, BeiDou and QZSS to the IGS MGEX [34.101, 102].
- Centre National d'Etudes Spatiales (CNES, French space agency) develops the GINS/DYNAMO software for GNSS POD. It is used by the CNES IGS AC to produce GPS, GLONASS, and Galileo POD solutions on a routine basis. The software is also used for LEO POD processing [34.103].
- The Position and Navigation Data Analysis (PANDA) software developed at Wuhan University, China, is another software package for GNSS and LEO POD. Wuhan University contributes multi-GNSS products to the IGS MGEX project as well as GPS products to the IGS Rapid combination (currently in evaluation mode) [34.104].
- GAMIT-GLOBK is a GPS processing software developed at the Department of Earth Atmospheric and Planetary Sciences at the Massachusetts Institute of Technology (MIT) [34.105], and is used by the MIT IGS AC to contribute weekly final as well as reprocessed products to the IGS.
- PAGES (Program for the Adjustment of GPS EphemerideS) is developed by the U.S. National Geodetic Survey (NGS) [34.106] and is used to produce GPS orbits, station parameters, and EOPs using a double-difference approach. The NGS is an IGS AC contributing ultra-rapid, rapid, final, and reprocessed products.
- A number of commercial services produce precise GNSS POD solutions using both in-house and externally developed software packages. These providers focus primarily on real-time or low-latency postprocessed products of interest to customers operating in areas such as precise marine and land navigation, cellular device positioning, GNSS integrity monitoring, and meteorological analysis. Providers active in this space include John Deere (Navcom) [34.107], JPL Global Differential GNSS System [34.108], Fugro [34.109], RX Networks [34.110], Trimble [34.111], and Veripos [34.112]. These providers in some cases manage their own ground station networks, operate processing centers, provide real-time GNSS orbits and clocks (typically transmitted as corrections to the broadcast ephemeris) and integrity data via geostationary satellite links, and even sell proprietary receiver hardware. GPS and GLONASS products are standard, with BeiDou and Galileo solutions quickly coming online.

34.6 Products

Precise GNSS orbit and clock products must include several items. There are of course the transmitter orbits: these conventionally refer to the CoM and are given as time series of ECEF coordinates and optionally velocities. Given the satellites' altitude, the orbits are sufficiently dynamic to allow accurate interpolation of coordinates given at 15 min intervals (or less) using an 8–11th order interpolator ([34.113] and Annex A.2.1). If desired, the resulting satellite positions can be expressed in an ECI using the EOPs provided along with the orbit and clock products.

Transmitter clock offsets may be provided at various intervals. Solutions submitted to the IGS generally employ intervals of 5 min or 30 s, while intervals as small as 1 s are common in real-time systems. In general, the smaller the interval the better, since a user must interpolate clocks to epochs falling between estimates. Interpolation introduces some error depending on how well the interpolation (typically piecewise-linear) fits the true clock offsets.

It is important to note that clock offsets as provided in the IGS products refer to the transmitter

antenna phase center while the orbits refer to the CoM. Since GNSS range data represent the geometric distance between the electrical phase center of the transmitter and receiver antennas, a user must adjust the CoM location given in the orbit product to the phase center. The products should therefore provide information about the phase center model (PCOs and PCVs) used so that a user can apply a consistent model. Metadata, for instance the version of IERS conventions, and the terrestrial reference frame realized in a set of products, is also useful so users can apply consistent models in their processing. This information is provided in the analysis strategy summary files [34.114].

IGS products are commonly named according to latency, including real-time (seconds), ultra-rapid (hours), rapid (next day), final (1–2 weeks), and reprocessed (every few years). Accuracy improves as latency increases for several reasons. Waiting longer tends to increase the available tracking data, especially for outlying stations, improving the distribution of the tracking network. Waiting allows one to use

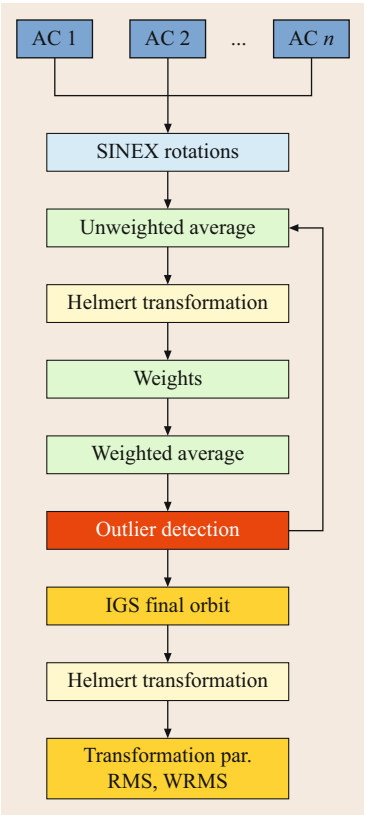


Fig. 34.8 Flow chart of the IGS orbit combination

ionosphere also improve as they progress from predicted quantities to estimates based on observations. The chief benefits of reprocessing campaigns are to have all tracking data available and to apply consistent models and estimation strategies over a long time span.

34.6.1 IGS Orbit and Clock Combination

The official IGS orbit and clock products are the result of a combination of the contributions of the individual ACs. Three product lines with different latency and accuracy are provided (Sect. 33.3):

- Ultra-rapid (observed half: 3–9 h, 3 cm, 50 ps)
- Rapid (17–41 h, 2.5 cm, 25 ps)
- Final (12–18 d, 2 cm, 20 ps).

The IGS Analysis Center Coordinator (ACC) is responsible for the generation of the combined products. A combined orbit product provides higher reliability and precision compared to the individual AC orbits. In the following, only the generation of the final orbit and clock products is discussed. The general combination methodology is described in [34.115].

Orbit Combination

Input for the orbit combination are satellite positions provided by the IGS ACs in an ECEF reference frame in SP3 format (Annex A.2.1) at 15 min sampling. The combination is based on an iterative weighted averaging (Fig. 34.8). To guarantee consistency of the station coordinates, EOPs, and orbits, AC-specific rotations are applied to the orbits prior to the actual combination [34.116]. These rotations have previously been derived by the IGS reference frame coordinator based on an analysis and combination of the station coordinates and EOPs that have been delivered by the ACs in the so-called SINEX (Solution Independent Exchange) format

more accurate nominal orbits and clocks, or provides time to iterate upon solutions to create improved nominals. This particularly benefits the editing of tracking data in preprocessing. One can use raw measurements or linear combinations thereof to detect unreasonable data and carrier-phase cycle slips, but the use of accurate orbit and clocks for data editing greatly enhances ones ability to screen for outlier tracking data. Other nominal models such as EOPs, zenith troposphere and associated mapping functions, and second-order

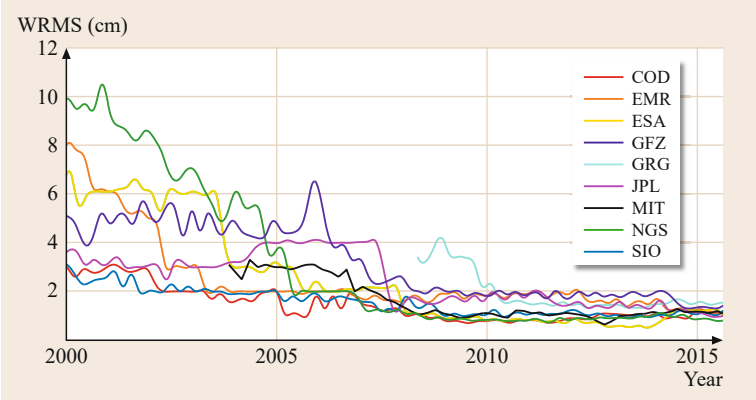


Fig. 34.9 Smoothed WRMS of individual AC GPS orbit solutions w.r.t. combined IGS final orbit

(Annex A.2.3). Detailed statistics as well as information on the combination are given in the weekly IGS combination summary files `igswww7.sum` (`www` stands for the GPS week), which are available at <ftp://ftp.cddis.eosdis.nasa.gov/pub/> in the corresponding `www` subdirectory.

The historical weighted root-mean square (WRMS) of the GPS orbits of the individual ACs w.r.t. the combined IGS final orbit is shown in Fig. 34.9. The comparison of the orbits improves with time due to the application of more sophisticated and consistent models and processing techniques. For example, in August 2007 JPL adopted the `igs05.atx` antenna model [34.117] already used at the time by the other ACs, resulting in a WRMS decrease by a factor of about 3. Starting with 2008, two groups of ACs can be distinguished: one group agreeing at the 2 cm level (EMR, GRG, JPL, SIO) and another group agreeing at the 1 cm level (COD, ESA, GFZ, MIT, NGS). In April 2014, ESA started to use an a priori box-wing model [34.118]. Although this model improves the orbit quality, it introduces larger differences w.r.t. the other AC solutions [34.119], resulting in a higher WRMS and a lower weight in the combination. As a consequence, the general WRMS of the individual IGS ACs approaches the 1.5 cm level. Further details on the precision of IGS final orbits are given in [34.120]. The combined IGS final GLONASS orbits are generated with the same procedure but in a separate process.

The WRMS of the orbit combination is, however, only a measure of the internal consistency of the orbits, which may suffer from common systematic errors. Harmonics of the draconitic GPS year (time period between the same orientation of the orbital planes w.r.t. the Sun, ≈ 351 d for GPS) have been reported for almost all IGS products [34.121]. It was shown that more sophisticated orbit modeling with an adjustable box-wing SRP model can reduce these draconitic errors [34.122]. However, deficiencies in the subdaily EOP model are also identified in [34.121] as causing artificial periodicities around 7, 9, 14, and 29 d due to aliasing.

The optical SLR technique allows for independent validation of the GNSS satellite orbits determined from microwave observations. An SLR analysis of 20 years of GPS and 12 years of GLONASS orbits computed by the CODE AC was performed by [34.123]. They found a 1 cm bias and 2 cm RMS for the two GPS satellites equipped with laser retro-reflectors. Whereas the SLR bias of the GLONASS satellites is in general on the few millimeters level, the RMS is around 3–4 cm. These numbers illustrate the discrepancies between internal precision as represented by the orbit combination WRMS and the accuracy as evaluated by SLR, which are caused by the systematic errors mentioned above.

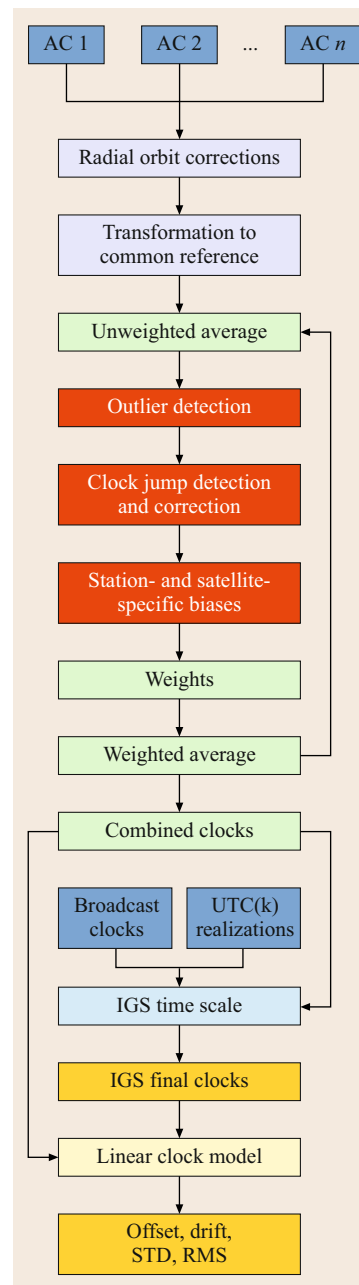


Fig. 34.10 Flow chart of the IGS clock combination. UTC(k) refers to UTC realizations at dedicated timing laboratories with calibrated GPS receivers included in the combined clocks

Clock Combination

Input for the clock combination are satellite and receiver clock estimates in RINEX clock format (Annex A.2.2). The general combination procedure is illustrated in Fig. 34.10 and discussed in more detail in [34.124, 125]. As a first step, radial orbit differences between the combined and individual AC orbits are computed and applied to the clocks in order to remove orbit-related systematic errors (radial or-

bit differences and satellite clock offsets are one-to-one correlated). Then, the individual AC clocks are aligned w.r.t. a common reference. Clock offset and drift w.r.t. a selected reference are removed for all ACs. The reference is taken from either the broadcast clock corrections or the clock estimates of a selected AC aligned to the broadcast clock corrections in a previous step.

During the iterative combination process, outliers are detected and clock jumps are corrected. A weighted average is formed based on weights determined from the deviation of the AC clocks w.r.t. an unweighted mean. The combined clocks are used to realize the IGS timescale (IGST, [34.126]). IGST is aligned to UTC via calibrated GPS receivers at time laboratories (labeled UTC(k) in Fig. 34.10) and GPS time from the navigation message. As a final step, the clock summary files are generated providing offset/drift of a linear clock model and RMS/STD w.r.t. the combined clocks as well as information about the time scale generation.

The historical RMS of the AC-specific GPS clock solutions w.r.t. the combined IGS final clocks is illustrated in Fig. 34.11. SIO does not provide clock corrections and NGS is excluded from the combination as only broadcast clocks are provided (due to double-difference observable processing). The RMS of the most consistent ACs is on the 100 ps level. For GLONASS, no combined clock product is available as mentioned in Sect. 34.4.8.

A critical issue for the clock combination is the consistency of the applied transmitter antenna and attitude models. As an example, all IGS ACs except for JPL switched the antenna model from *igs05.atx* to *igs08.atx* in April 2011 [34.127]. For JPL, this switch took place in July 2011 [34.117]. The 3 months of inconsistent antenna modeling can be clearly seen in Fig. 34.11 as JPL's clock RMS increases by a factor of more than 3. Discrepancies among ACs in attitude modeling are evident for the eclipse periods of the GPS Block II/IIA

satellites. These cause large clock differences responsible for the rejection of individual ACs during the combination process [34.128]. Figure 34.12 illustrates this problem for GPS Block IIA SVN-33. The MIT clock estimates show a significantly different behavior after the satellite leaves the shadow, resulting in an exclusion of this AC for this particular satellite.

34.6.2 Formats and Transmission

A variety of GNSS product formats and transmission mechanisms are used today (Annex A). Postprocessed products are usually provided as compressed files. Open standards for data exchange include SP3 (standard product 3) for GNSS orbits and clocks [34.129], clock RINEX for receiver and transmitter clocks [34.130], Earth rotation parameter (ERP) [34.131], and antenna exchange (ANTEX) antenna calibration [34.132] formats. Within the IGS, the fourth comment line in the SP3 orbit files is used to document important modeling options including the phase center model, name of the ocean tidal loading and atmospheric tidal loading models, and whether CoM corrections are applied. The full format description of this SP3 comment line is given in [34.133]. Each POD software generally also uses proprietary formats representing the same information, for example, the GIPSY *pos* or Bernese *standard orbit* [34.68] formats.

Real-time systems distribute products via low-latency files (e.g., each minute) as well as few second latency streams. The real-time streams usually represent the precise orbit and clock solution as corrections to the current broadcast navigation message (with respect to a specific issue of data encoded in the corrections message). This implies that the generator of the corrections must account for any differences in the transmitter antenna phase center offsets used to compute the broadcast ephemerides and precise solutions. The user applies the corrections to the broadcast or-

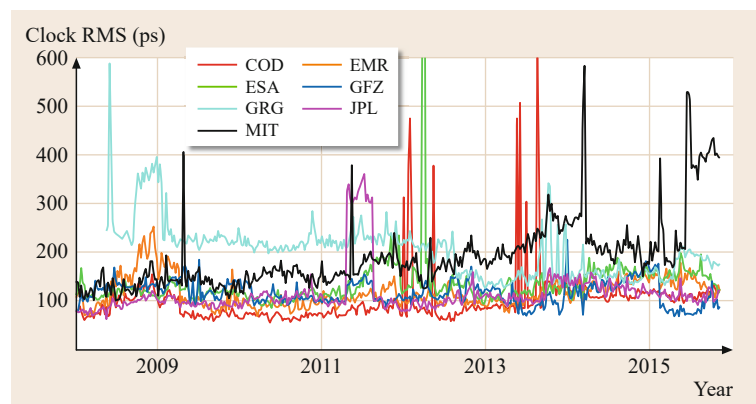


Fig. 34.11 RMS of individual AC clock solutions w.r.t. combined IGS final clocks

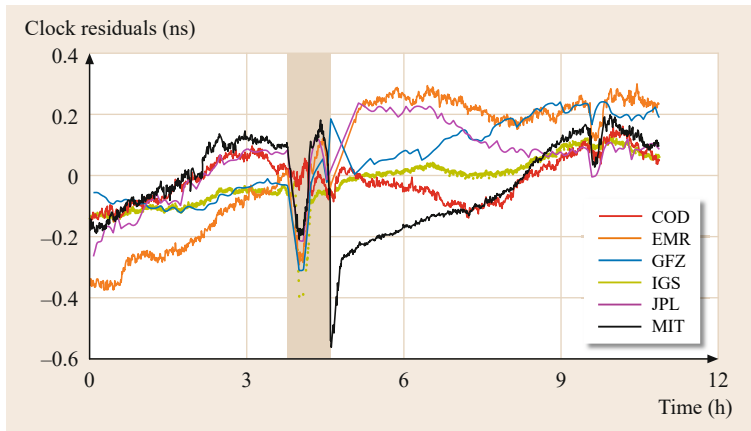


Fig. 34.12 Individual AC and combined IGS clock residuals of the GPS Block IIA satellite SVN-33 for 6 February 2011. The *light brown shaded area* indicates the eclipse period. The ESA clock estimates are used as reference clock and offset/drift of each AC are removed. EMR and MIT clock estimates of this satellite were excluded from the clock combination on that day

bits and clocks directly. The corrections approach has the benefit that some latency (up to tens of seconds) does not significantly degrade the user solution since the navigation message accounts for the majority of the bias and drift in the satellite position and clock, whereas the corrections terms are relatively steady over short time intervals (at least for nominally performing transmitter clocks). Corrections are usually transmitted at 1 s intervals, and losing some corrections over the communication link, or updating at longer intervals, is possible with this approach. Real-time correction streams from commercial providers are encoded in proprietary binary formats sent as TCP or UDP packets over the Internet or transmitted to user equipment via geostationary satellite links [34.134]. The IGS real-time service transmits corrections over the Internet in the open State-Space Representation (SSR) format [34.135].

34.6.3 Using Products

Whether products are acquired as postprocessed files or real-time streams, the user applies a set of precise orbit, clock, and ancillary information in their processing. Maintaining consistency with GNSS products is

critical to realizing the best possible accuracy. As discussed, orbits need to be adjusted from the CoM to the antenna phase center, which requires knowledge of both the antenna calibrations and the spacecraft attitude model used to generate the GNSS products. Attitude models in the POD software packages have varying levels of complexity, particularly for non-nominal attitude regimes (e.g., eclipse). The best consistency is therefore achieved by using the same software that generated the products.

The data type to which the clocks products refer should also be consistent. By convention, GPS (broadcast ephemeris and precise) products provide clocks estimated using the L1 P(Y)/L2 P(Y) ionosphere-free linear combination, while GLONASS processing may use either the coarse or precise ranging code to form the ionosphere-free linear combination (it does not matter since the user needs to estimate range biases on each receiver-transmitter link again). For the Galileo and BeiDou CDMA systems conventions are currently developing (Sect. 34.4.8). In any case, the user should process the same data type used to generate the GNSS products. If the data type is not available, the user must apply the appropriate DCBs (e.g., GPS L1 C/A vs. L1 P(Y)) in preprocessing.

34.7 Outlook

For many years the POD community mainly focused on GPS since this was the only stable constellation. Over the course of nearly three decades, the knowledge of physical properties underlying the measurement system have been continually refined. The high accuracies achieved today are enabled by several key factors: taking advantage of tracking data provided by a large,

global set of geodetic stations, careful treatment of measurement biases, robust data editing schemes, sophisticated modeling of station motion, atmospheric effects, clock offsets, electrical phase centers, spacecraft dynamics and attitude.

In recent years, the renewal of the GLONASS constellation and the building of Galileo and BeiDou has

resulted in a total of four usable GNSSs. The level of knowledge about the newer systems, and accuracy of POD products, is in many ways comparable to the first decade of precision GPS. The challenges that lie ahead are multifaceted: areas such as the handling of measurement biases between the GNSSs, treatment of more complex signal structures, multi-GNSS ambiguity resolution, and the generation of consistent time scales provide rich grounds for research and development. Details regarding spacecraft attitude control, physical satellite properties and force models, antenna-phase centers, and constellation operations are still in limited distribution. It is, therefore, paramount for the precision GNSS community to engage system operators to make available detailed system information enabling precise POD. For the foreseeable future, GPS will remain the cornerstone of multi-GNSS processing, but the accuracies achieved with the other systems should rapidly improve.

For all GNSSs, model improvements and the determination of clock offsets remain important research topics. Orbit parameters are constrained by well-understood dynamics, but clocks are treated as unconstrained epoch parameters to mitigate for steering, reset events, and other difficult-to-model behaviors. Epoch-independent parameterization, coupled with least-squares estimation, allows clock estimates to absorb model errors to minimize postfit residuals.

Since clock estimates refer to antenna phase centers, accurate models for phase and group delay (receiver and transmitter) and spacecraft attitude are particularly critical. Correlations between receiver clock, geodetic station height, and zenith tropospheric delay are concerns for some science applications. Highly stable satellite clocks may alleviate some of these issues. For instance, the Galileo passive hydrogen masers allow for modeling the clock instead of epoch-wise estimation. It was demonstrated that this approach promises to also improve the orbit quality [34.136].

A challenge for the IGS is the development and implementation of a new software for a fully consistent combination of orbits and clocks of multiple GNSSs, also known as *ACC 2.0*. Currently, combined GPS and GLONASS products are generated by completely separate processing chains. Furthermore, no combined Galileo and BeiDou products are generated although several ACs provide solutions for these systems. The need for such a software upgrade has long been recognized but little progress has been made so far. Related to this, the determination and communication of GNSS timescales in a multi-GNSS world requires an expanded set of signal conventions.

Increasing demands from scientific users drive many of these challenges. GNSS provides important baseline and orientation information to the determination of the global terrestrial reference frame, but does not contribute to geocenter and scale due to the lack of independent, absolute antenna calibrations. Many precise geodetic and atmospheric science applications are taking advantage of the improved spatial and temporal observation coverage offered by multiple GNSS constellations, but continue to observe signals in physical parameters at frequencies related to GNSS spacecraft dynamic forces. The societal benefits provided by emerging low-latency, high-accuracy GNSS applications in areas such as tsunami and earthquake early warning place significant demands on GNSS measurement and processing infrastructure. These are only some examples illustrating the continued promise of precise GNSS POD and its applications, with technical and scientific rewards that will progress well beyond this generation.

Acknowledgments. We would like to thank the Editors for their helpful reviews during the development of this chapter and Mathias Fritsche, Deutsches Geo-ForschungsZentrum Potsdam (GFZ) for providing information on orbit and clock combination.

References

- 34.1 O. Montenbruck, P. Steigenberger, A. Hauschild: Broadcast versus precise ephemerides: A multi-GNSS perspective, *GPS Solut.* **19**(2), 321–333 (2015)
- 34.2 G. Petit, B. Luzum: *IGRS Conventions (2010)*, IERS Technical Note No. 36 (Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt a. M. 2010)
- 34.3 M. Rothacher, G. Beutler, T.A. Herring, R. Weber: Estimation of nutation using the Global Positioning System, *J. Geophys. Res.* **104**(B3), 4835–4859 (1999)
- 34.4 B.J. Luzum, J.R. Ray, M.S. Carter, F.J. Josties: Recent improvements to IERS Bulletin A combination and prediction, *GPS Solut.* **4**(3), 34–40 (2001)
- 34.5 C. Bizouard, D. Gambis: The combined solution C04 for Earth Orientation Parameters consistent with International Terrestrial Reference Frame 2008, Observatoire de Paris, <https://hpiers.obspm.fr/iers/eop/eopc04/C04.guide.pdf>
- 34.6 P. Rebischung, J. Griffiths, J. Ray, R. Schmid, X. Collilieux, B. Garayt: IGS08: the IGS realization of ITRF2008, *GPS Solut.* **16**(4), 483–494 (2012)
- 34.7 P.M. Mathews, V. Dehant, J.M. Gipson: Tidal station displacements, *J. Geophys. Res.* **102**(B9), 20469–20477 (1997)
- 34.8 H.-G. Scherneck: A parametrized solid Earth tide model and ocean loading effects for

- global geodetic base-line measurements, *Geophys. J. Int.* **106**(3), 677–694 (1991)
- 34.9 F. Lyard, F. Lefevre, T. Letellier, O. Francis: Modelling the global ocean tides: Modern insights from FES2004, *Ocean Dyn.* **56**(5/6), 394–415 (2006)
- 34.10 L. Carrere, F. Lyard, A. Guillot, M. Cancet: FES 2012: A new tidal model taking advantage of nearly 20 years of altimetry measurements, *Proc. 20 Years Prog. Radar Altimetry Symp.*, Venice-Lido (CNES/ESA, Toulouse 2012) p. 5
- 34.11 S.D. Desai: Observing the pole tide with satellite altimetry, *J. Geophys. Res.* **107**(C11,3180), 1–13 (2003) doi:[10.1029/2001JC001224](https://doi.org/10.1029/2001JC001224)
- 34.12 R.D. Ray, R.M. Ponte: Barometric tides from ECMWF operational analyses, *Ann. Geophys.* **21**(8), 1897–1910 (2003)
- 34.13 J. Boehm, R. Heinkelmann, H. Schuh: Short Note: A global model of pressure and temperature for geodetic applications, *J. Geod.* **81**(10), 679–683 (2007)
- 34.14 J. Boehm, A. Niell, P. Tregoning, H. Schuh: Global Mapping Function (GMF): A new empirical mapping function based on numerical weather model data, *Geophys. Res. Lett.* **33**(L07304), 1–4 (2006) doi:[10.1029/2005GL025546](https://doi.org/10.1029/2005GL025546)
- 34.15 J. Böhm, G. Möller, M. Schindelegger, G. Pain, R. Weber: Development of an improved empirical model for slant delays in the troposphere (GPT2w), *GPS Solut.* **19**(3), 433–441 (2014)
- 34.16 J. Boehm, B. Werl, H. Schuh: Troposphere mapping functions for GPS and very long baseline interferometry from European Centre for Medium-Range Weather Forecasts operational analysis data, *J. Geophys. Res.* **111**(B02406), 1–9 (2006) doi:[10.1029/2005JB003629](https://doi.org/10.1029/2005JB003629)
- 34.17 M. Fritsche, R. Dietrich, C. Knöfel, A. Rülke, S. Vey, M. Rothacher, P. Steigenberger: Impact of higher-order ionospheric terms on GPS estimates, *Geophys. Res. Lett.* **32**(L23311), 1–5 (2005) doi:[10.1029/2005GL024342](https://doi.org/10.1029/2005GL024342)
- 34.18 N. Ashby: Relativity in the global positioning system, *Living Rev.* **6**(1), 1–42 (2003) doi:[10.12942/lrr-2003-1](https://doi.org/10.12942/lrr-2003-1)
- 34.19 R. Schmid, R. Dach, X. Collilieux, A. Jäggi, M. Schmitz, F. Dilssner: Absolute IGS antenna phase center model igs08.atx: Status and potential improvements, *J. Geod.* **90**(4), 343–364 (2015)
- 34.20 J.T. Wu, S.C. Wu, G.G. Hajj, W.I. Bertiger, S.M. Lichten: Effects of antenna orientation on GPS carrier-phase, *Manuscr. Geod.* **18**, 91–98 (1993)
- 34.21 O. Montenbruck, R. Schmid, F. Mercier, P. Steigenberger, C. Noll, R. Fatkulin, S. Kogure, A.S. Ganeshan: GNSS satellite geometry and attitude models, *Adv. Space Res.* **56**(6), 1015–1029 (2015)
- 34.22 J. Kouba: A simplified yaw-attitude model for eclipsing GPS satellites, *GPS Solut.* **13**(1), 1–12 (2009)
- 34.23 F. Dilssner: GPS IIF-1 satellite, antenna phase center and attitude modeling, *Inside GNSS* **5**(6), 59–64 (2010)
- 34.24 F. Dilssner, T. Springer, G. Gienger, J. Dow: The GLONASS-M satellite yaw-attitude model, *Adv. Space Res.* **47**(1), 160–171 (2011)
- 34.25 X. Dai, M. Ge, Y. Lou, C. Shi, J. Wickert, H. Schuh: Estimating the yaw-attitude of BDS IGS0 and ME0 satellites, *J. Geod.* **89**(10), 1005–1018 (2015)
- 34.26 Y. Ishijima, N. Inaba, A. Matsumoto, K. Terada, H. Yonechi, H. Ebisutani, S. Ukawa, T. Okamoto: Design and development of the first Quasi-Zenith Satellite attitude and orbit control system, *IEEE Aerosp. Conf.* (2009) doi:[10.1109/AERO.2009.4839537](https://doi.org/10.1109/AERO.2009.4839537)
- 34.27 C.J. Rodriguez-Solano: Impact of Albedo Modelling on GPS Orbits, Master Thesis (TU München, Munich 2009)
- 34.28 M. Ziebart, S. Edwards, S. Adhya, A. Sibthorpe, P. Arrowsmith, P. Cross: High precision GPS IIR orbit prediction using analytical non-conservative force models, *Proc. ION GNSS 2004*, Long Beach (ION, Virginia 2004) pp. 1764–1770
- 34.29 IGS: GPS transmit power levels, <http://acc.igs.org/orbits/thrust-power.txt>
- 34.30 N.K. Pavlis, S.A. Holmes, S.C. Kenyon, J.K. Factor: The development and evaluation of the Earth Gravitational Model 2008 (EGM2008), *J. Geophys. Res.* **117**(B04406), 1–38 (2012) doi:[10.1029/2011JB008916](https://doi.org/10.1029/2011JB008916)
- 34.31 C.C. Finlay, S. Maus, C.D. Beggan, T.N. Bondar, A. Chambodut, T.A. Chernova, A. Chuliat, V.P. Golovkov, B. Hamilton, M. Hamoudi, R. Holme, G. Hulot, W. Kuang, B. Langlais, V. Lesur, F.J. Lowes, H. Luhr, S. Macmillan, M. Manda, S. McLean, C. Manoj, M. Menvielle, I. Michaelis, N. Olsen, J. Rauberg, M. Rother, T.J. Sabaka, A. Tangborn, L. Toffner-Clausen, E. Thebault, A.W.P. Thomson, I. Wardinski, Z. Wei, T.I. Zvereva: International Geomagnetic Reference Field: The eleventh generation, *Geophys. J. Int.* **183**(3), 1216–1230 (2010)
- 34.32 Chalmers University: *Online Ocean Tide Loading Computation Service* <http://holt.oso.chalmers.se/loading>
- 34.33 L. Petrov, J.-P. Boy: Study of the atmospheric pressure loading signal in very long baseline interferometry observations, *J. Geophys. Res.* **109**(B03405), 1–14 (2004) doi:[10.1029/2003JB002500](https://doi.org/10.1029/2003JB002500)
- 34.34 P. Tregoning, C. Watson, G. Ramillien, H. McQueen, J. Zhang: Detecting hydrologic deformation using GRACE and GPS, *Geophys. Res. Lett.* **36**(L1540), 1–6 (2009) doi:[10.1029/2009GL038718](https://doi.org/10.1029/2009GL038718)
- 34.35 P. Misra, P. Enge: *Global Positioning System Signals, Measurements, and Performance*, 2nd edn. (Ganga-Jamuna, Lincoln 2006)
- 34.36 J. Boehm, H. Schuh: Vienna mapping functions in VLBI analyses, *Geophys. Res. Lett.* **31**(L01603), 1–4 (2004) doi:[10.1029/2003GL018984](https://doi.org/10.1029/2003GL018984)
- 34.37 J. Saastamoinen: Atmospheric correction for the troposphere and stratosphere in radio ranging of satellites. In: *The Use of Artificial Satellites for Geodesy*, Geophysical Monograph Series, Vol. 15, ed. by S.W. Henriksen, A. Mancini, B.H. Chovitz

- (AGU, Washington 1972) pp. 247–251
- 34.38 J.L. Davis, T.A. Herring, I.I. Shapiro, A.E.E. Rogers, G. Elgered: Geodesy by radio interferometry: Effects of atmospheric modeling errors on estimates of baseline length, *Radio Sci.* **20**(6), 1593–1607 (1985)
- 34.39 D. Bilitza, D. Altadill, Y. Zhang, C. Mertens, V. Truhlik, P. Richards, L. McKinnell, B. Reinisch: The International Reference Ionosphere 2012 – A model of international collaboration, *J. Space Weather Space Clim.* **4**, A07 (2014)
- 34.40 S. Kedar, G.A. Hajj, B.D. Wilson, M.B. Heflin: The effect of the second order GPS ionospheric correction on receiver positions, *Geophys. Res. Lett.* **30**(16), 1–4 (2003) doi:[10.1029/2003GL017639](https://doi.org/10.1029/2003GL017639)
- 34.41 E.J. Petrie, M. Hernández-Pajares, P. Spalla, P. Moore, M.A. King: A review of higher order ionospheric refraction effects on dual frequency GPS, *Surv. Geophys.* **32**(3), 197–253 (2011)
- 34.42 M. Garcia-Fernandez, S.D. Desai, M.D. Butala, A. Komjathy: Evaluation of different approaches to modeling the second-order ionospheric delay on GPS measurements, *J. Geophys. Res. Space Phys.* **118**(12), 7864–7873 (2013)
- 34.43 M. Hernández-Pajares, J.M. Juan, J. Sanz, R. Orús: Second-order ionospheric term in GPS: Implementation and impact on geodetic estimates, *J. Geophys. Res.* **112**(B08417), 1–16 (2007) doi:[10.1029/2006JB004707](https://doi.org/10.1029/2006JB004707)
- 34.44 S. Bassiri, G.A. Hajj: Higher-order ionospheric effects on the global positioning system observables and means of modeling them, *Manuscr. Geod.* **18**, 280–289 (1993)
- 34.45 N. Ashby, J.J. Spilker Jr.: Introduction to relativistic effects on the Global Positioning System. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker Jr. (AIAA, Washington 1996) pp. 623–697
- 34.46 J. Kouba: Relativistic time transformations in GPS, *GPS Solut.* **5**(4), 1–9 (2002)
- 34.47 R. Schmid, P. Steigenberger, G. Gendt, M. Ge, M. Rothacher: Generation of a consistent absolute phase center correction model for GPS receiver and satellite antennas, *J. Geod.* **81**(12), 781–798 (2007)
- 34.48 Y.E. Bar-Sever: A new model for GPS yaw attitude, *J. Geod.* **70**(11), 714–723 (1996)
- 34.49 Y. Bar-Sever, D. Kuang: New empirically derived solar radiation pressure model for Global Positioning System satellites, *IPN Prog. Rep.* **42**, 159 (2004)
- 34.50 J.P. Weiss, Y. Bar-Sever, W. Bertiger, S. Desai, M. Garcia-Fernandez, B. Haines, D. Kuang, C. Selle, A. Sibois, A. Sibthorpe: Orbit and attitude modeling at the JPL Analysis Center, Int. GNSS Serv. Workshop, Pasadena (IGS, Pasadena 2014)
- 34.51 H.F. Fliegel, T.E. Gallini: Solar force modeling of Block IIR Global Positioning System satellites, *J. Spacecr. Rockets* **33**(6), 863–866 (1996)
- 34.52 M. Ziebart, S. Adhya, A. Sibthorpe, S. Edwards, P. Cross: Combined radiation pressure and thermal modelling of complex satellites: Algorithms and on-orbit tests, *Adv. Space Res.* **36**(3), 424–430 (2005)
- 34.53 P.C. Knocke, J.C. Ries, B.D. Tapley: Earth radiation pressure effects on satellites, *Proc. AIAA/AAS Astrodyn. Conf.*, Minneapolis (AIAA, Reston 1988) pp. 577–587
- 34.54 M. Ziebart, A. Sibthorpe, P. Cross, Y. Bar-Sever, B. Haines: Cracking the GPS-SLR orbit anomaly, *Proc. ION GNSS 2007*, Fort Worth (ION, Virginia 2007) pp. 2033–2038
- 34.55 C.J. Rodriguez-Solano, U. Hugentobler, P. Steigenberger, S. Lutz: Impact of Earth radiation pressure on GPS position estimates, *J. Geod.* **86**(5), 309–317 (2012)
- 34.56 B.A. Wielicki, B.R. Barkstrom, E.F. Harrison, R.B. Lee, G.L. Smith, J.E. Cooper: Clouds and the Earth's radiant energy system (CERES): An Earth observing system experiment, *Bull. Am. Meteorol. Soc.* **77**(5), 853–868 (1996)
- 34.57 United States Coast Guard: <https://www.navcen.uscg.gov/?Do=constellationstatus>
- 34.58 United States Coast Guard: <https://www.navcen.uscg.gov/?pageName=currentNanUS>
- 34.59 Information and Analysis Center for Positioning, Navigation and Timing: <https://www.glonass-iac.ru/en/GLONASS>
- 34.60 Information and Analysis Center for Positioning, Navigation and Timing: <https://www.glonass-iac.ru/en/CUSGLONASS/>
- 34.61 European GNSS Service Centre: <http://www.gsc-europa.eu/system-status/Constellation-Information>
- 34.62 European GNSS Service Centre: <http://www.gsc-europa.eu/system-status/user-notifications>
- 34.63 Cabinet Office: <http://qzss.go.jp/en/technical/satellites/index.html#QZSS>
- 34.64 JAXA: <http://qz-vision.jaxa.jp/USE/en/naqu>
- 34.65 G.J. Bierman: *Factorization Methods for Discrete Sequential Estimation* (Academic Press, New York 1977)
- 34.66 P. Axelrad, R.G. Brown: GPS navigation algorithms. In: *Global Positioning System: Theory and Applications*, Vol. 1, ed. by B.W. Parkinson, J.J. Spilker Jr. (AIAA, Washington 1996) pp. 409–433
- 34.67 B. Tapley, B. Schutz, G.H. Born: *Statistical Orbit Determination* (Academic Press, Burlington 2004)
- 34.68 R. Dach, F. Andritsch, D. Arnold, S. Bertone, P. Fridez, A. Jäggi, Y. Jean, A. Maier, L. Mervart, U. Meyer, E. Orliac, E. Ortiz-Geist, L. Prange, S. Scaramuzza, S. Schaer, D. Sidorov, A. Sušnik, A. Villiger, P. Walser, C. Baumann, G. Beutler, H. Bock, A. Gäde, S. Lutz, M. Meindl, L. Ostini, K. Sošnica, A. Steinbach, D. Thaller: *Bernese GNSS Software Version 5.2*, ed. by R. Dach, S. Lutz, P. Walser, P. Fridez (Astronomical Institute, University of Bern, Bern 2015)
- 34.69 M. Rothacher: Estimation of station heights with GPS. In: *Vertical Reference Systems, International Association of Geodesy Symposia*, Vol. 124, ed. by H. Drewes, A.H. Dodson, P.S. Fortes, L. Sanchez, P. Sandoval (Springer, Berlin, Heidelberg 2002)

- pp. 81–90
- 34.70 S. Jin, J. Wang, P.-H. Park: An improvement of GPS height estimations: Stochastic modeling, *Earth Planets Space* **57**(4), 253–259 (2014)
- 34.71 X. Luo, M. Mayer, B. Heck, J.L. Awange: A realistic and easy-to-implement weighting model for GPS phase observations, *IEEE Trans. Geosci. Remote Sens.* **52**(10), 6110–6118 (2014)
- 34.72 D.S. MacMillan, C. Ma: Atmospheric gradients and the VLBI terrestrial and celestial reference frames, *Geophys. Res. Lett.* **24**(4), 453–456 (1997)
- 34.73 O. Titov, V. Tesmer, J. Boehm: OCCAM v. 6.0 software for VLBI data analysis, *Proc. IVS 2004 Gen. Meet.* (2004) pp. 267–271
- 34.74 S. Wu, T.P. Yuncck, C.L. Thornton: Reduced-dynamic technique for precise orbit determination of low Earth satellites, *J. Guid. Control Dyn.* **14**(1), 24–30 (1991)
- 34.75 G. Beutler, E. Brockmann, W. Gurtner, U. Hugentobler, L. Mervart, M. Rothacher, A. Verdun: Extended orbit modeling techniques at the CODE processing center of the international GPS service for geodynamics (IGS): Theory and initial results, *Manuscr. Geod.* **19**, 367–386 (1994)
- 34.76 T.A. Springer, G. Beutler, M. Rothacher: A new solar radiation pressure model for GPS satellites, *GPS Solut.* **2**(3), 50–62 (1999)
- 34.77 D. Arnold, M. Meindl, G. Beutler, R. Dach, S. Schaer, S. Lutz, L. Prange, K. Soñnica, L. Mervart, A. Jäggi: CODE's new solar radiation pressure model for GNSS orbit determination, *J. Geod.* **89**(8), 775–791 (2015)
- 34.78 A. Sibthorpe, W. Bertiger, S.D. Desai, B. Haines, N. Harvey, J.P. Weiss: An evaluation of solar radiation pressure strategies for the GPS constellation, *J. Geod.* **85**(8), 505–517 (2011)
- 34.79 N. Romero: CC2NONCC update to handle more than 24 satellites per epoch, IGSMAIL-6542 (2012) <https://igsb.jpl.nasa.gov/pipermail/igsbmail/2012/007732.html>
- 34.80 P. Steigenberger, O. Montenbruck, U. Hessels: Performance evaluation of the early CNAV navigation message, *Navigation* **62**(3), 219–228 (2015)
- 34.81 O. Montenbruck, A. Hauschild, P. Steigenberger: Differential code bias estimation using multi-GNSS observations and global ionosphere maps, *Navigation* **61**(3), 191–201 (2014)
- 34.82 L. Mervart: Ambiguity Resolution Techniques in Geodetic and Geodynamic Applications of the Global Positioning System, Ph.D. Thesis, Geodätisch-geophysikalische Arbeiten in der Schweiz, Vol. 53 (Schweizerische Geodätische Kommission, Zürich 1995)
- 34.83 M. Ge, G. Gendt, G. Dick, F.P. Zhang: Improving carrier-phase ambiguity resolution in global GPS network solutions, *J. Geod.* **79**(1), 103–110 (2005)
- 34.84 S. Loyer, F. Perosanz, F. Mercier, H. Capdeville, J.-C. Marty: Zero-difference GPS ambiguity resolution at CNES-CLS IGS Analysis Center, *J. Geod.* **86**(11), 991–1003 (2012)
- 34.85 P. Steigenberger, U. Hugentobler, S. Loyer, F. Perosanz, L. Prange, R. Dach, M. Uhlemann, G. Gendt, O. Montenbruck: Galileo orbit and clock quality of the IGS multi-GNSS experiment, *Adv. Space Res.* **55**(1), 269–281 (2015)
- 34.86 O. Montenbruck, P. Steigenberger, U. Hugentobler: Enhanced solar radiation pressure modeling for Galileo satellites, *J. Geod.* **89**(3), 283–297 (2015)
- 34.87 P. Steigenberger, U. Hugentobler, A. Hauschild, O. Montenbruck: Orbit and clock analysis of Compass GEO and IGSO satellites, *J. Geod.* **87**(6), 515–525 (2013)
- 34.88 J. Liu, D. Gua, B. Ju, Z. Shen, Y. Lai, D. Yi: A new empirical solar radiation pressure model for BeiDou GEO satellites, *Adv. Space Res.* **57**(1), 234–244 (2016)
- 34.89 Russian Institute of Space Device Engineering: Global Navigation Satellite System GLONASS – Interface Control Document, v5.1, (Russian Institute of Space Device Engineering, Moscow 2008)
- 34.90 L. Wanninger: Carrier-phase inter-frequency biases of GLONASS receivers, *J. Geod.* **86**(2), 139–148 (2012)
- 34.91 International GNSS Service Analysis Center Coordinator, <http://acc.igs.org/>
- 34.92 Z. Deng, Q. Zhao, T. Springer, L. Prange, M. Uhlemann: Orbit and clock determination – BeiDou, *Proc. IGS Workshop, Pasadena* (IGS, Pasadena 2014)
- 34.93 P. Rebischung: IGB08, IGSMAIL-6663 (2012) <https://igsb.jpl.nasa.gov/pipermail/igsbmail/2012/006655.html>
- 34.94 X. Wu, J. Ray, T. van Dam: Geocenter motion and its geodetic and geophysical implications, *J. Geodyn.* **58**, 44–66 (2012)
- 34.95 R. Ferland, G. Gendt, T. Schöne: IGS reference frame maintenance, Celebrating a decade of the International GPS Service, Workshop and Symposium 2004, Bern, ed. by M. Meindl (Astronomical Institute, University of Bern, Bern 2005) pp. 13–34
- 34.96 CODE Analysis Strategy Summary (2016) <https://igsb.jpl.nasa.gov/igsb/center/analysis/code.acn>
- 34.97 A. Hauschild, O. Montenbruck: Real-time clock estimation for precise orbit determination of LEO-satellites, *Proc. ION GNSS 2008, Savannah* (ION, Virginia 2008) pp. 581–589
- 34.98 H. Bock, R. Dach, A. Jäggi, G. Beutler: High-rate GPS clock corrections from CODE: Support of 1 Hz applications, *J. Geod.* **83**(11), 1083–1094 (2009)
- 34.99 T.A. Springer: *NAPEOS Mathematical Models and Algorithms*, DOPS-SYS-TN-0100-OPS-GN (ESA/ESOC, Darmstadt 2009)
- 34.100 JPL: GIPSY-OASIS, <https://gipsy-oasis.jpl.nasa.gov>
- 34.101 G. Gendt, G. Dick, W. Soehne: GFZ analysis center of IGS – Annual report 1998. In: *IGS 1998 Technical Reports*, ed. by K. Goway, R. Neilan, A. Moore (JPL, Pasadena 1998) pp. 79–87
- 34.102 M. Ge, G. Gendt, G. Dick, F.P. Zhang, M. Rothacher: A new data processing strategy for huge GNSS global networks, *J. Geod.* **80**(4), 199–203 (2006)
- 34.103 J.C. Marty, S. Loyer, F. Perosanz, F. Mercier, G. Bracher, B. Legresy, L. Portier, H. Capdeville, F. Fund, J.M. Lemoine: GINS: The CNES/IGRS GNSS

- scientific software, Proc. 3rd Int. Coll. Sci. Fundam. Asp. Galileo Program., ESA WPP326, Copenhagen (ESA, Noordwijk 2011)
- 34.104 Q. Zhao, J. Guo, M. Li, L. Qu, Z. Hu, C. Shi, J. Liu: Initial results of precise orbit and clock determination for COMPASS navigation satellite system, *J. Geod.* **87**(5), 475–486 (2013)
- 34.105 MIT: GAMIT-GLOBK, <http://www-gpsg.mit.edu/~simon/gtgk/>
- 34.106 W.G. Kass, R.L. Dulaney, J. Griffiths, S. Hilla, J. Ray, J. Rohde: Global GPS data analysis at the National Geodetic Survey, *J. Geod.* **83**(3/4), 289–295 (2009)
- 34.107 K. Dixon: StarFire: A global SBAS for sub-decimeter precise point positioning, Proc. ION GNSS 2006, Fort Worth (ION, Virginia 2006) pp. 2286–2296
- 34.108 Global Differential GNSS System, <http://www.gdgps.net>
- 34.109 J. Tegeedor, D. Lapucha, O. Ørpen, E. Vigen, T. Melgard, R. Strandli: The new G4 service: Multi-constellation precise point positioning including GPS, GLONASS, Galileo and BeiDou, Proc. ION GNSS+ 2015, Tampa (ION, Virginia 2015) pp. 1089–1095
- 34.110 E. Derbez, R. Lee: GPStream: A low bandwidth architecture to deliver or autonomously generate predicted ephemeris, Proc. ION GNSS 2008, Savannah (ION, Virginia 2008) pp. 1258–1264
- 34.111 M. Glocker, H. Landau, R. Leandro, M. Nitschke: Global precise multi-GNSS positioning with Trimble Centerpoint RTX, Proc. 6th ESA Workshop Satell. Navig. Technol. Eur. Workshop GNSS Signals Signal Proces. (NAVITEC), Noordwijk (IEEE, New York 2012), doi:10.1109/NAVITEC.2012.6423060
- 34.112 C. Rocken, L. Mervart, J. Johnson, Z. Lukes, T. Springer, T. Iwabuchi, S. Cummins: A new real-time global GPS and GLONASS precise positioning correction service: Apex, Proc. ION GNSS 2011, Portland (ION, Virginia 2011) pp. 1825–1838
- 34.113 Y. Feng, Y. Zheng: Efficient interpolations to GPS orbits for precise wide area applications, *GPS Solut.* **9**(4), 273–282 (2005)
- 34.114 IGS Analysis Strategy Summaries (2016) <ftp://igs.org/pub/center/analysis>
- 34.115 G. Beutler, J. Kouba, T. Springer: Combining the orbits of the IGS analysis centers, *Bull. Geod.* **69**, 200–222 (1995)
- 34.116 J. Griffiths: Misalignment of the AC final orbits (2012) http://acc.igs.org/orbits/acc_report_final_rotations.pdf
- 34.117 S. Desai, W. Bertiger, B. Haines, D. Kuang, C. Selle, A. Sibois, A. Sibthorpe, J. Weiss: JPL IGS analysis center report, 2005–2012, Int. GNSS Serv. Techn. Rep. 2011, Pasadena, ed. by M. Meindl, R. Dach, Y. Jean (IGS Central Bureau, Pasadena 2012) pp. 85–90
- 34.118 C. Garcia Serrano, L. Agrotis, F. Dilssner, J. Feltens, M. van Kints, I. Romero, T. Springer, W. Enderle: The ESA/ESOC analysis center progress and improvements, IGS Workshop 2014, Pasadena (IGS Central Bureau, Pasadena 2014)
- 34.119 T. Springer, M. Otten, C. Flohrer, F. Pereira, F. Gini, W. Enderle: GNSS satellite orbit modeling at ESOC, IGS Workshop 2014, Pasadena (IGS Central Bureau, Pasadena 2014)
- 34.120 J. Griffiths, J. Ray: On the precision and accuracy of IGS orbits, *J. Geod.* **83**(3/4), 277–287 (2009)
- 34.121 J. Griffiths, J.R. Ray: Sub-daily alias and draconitic errors in the IGS orbits, *GPS Solut.* **17**(3), 413–422 (2012)
- 34.122 C.J. Rodríguez-Solano, U. Hugentobler, P. Steigenberger, M. Bloßfeld, M. Fritsche: Reducing the draconitic errors in GNSS geodetic products, *J. Geod.* **88**(6), 559–574 (2014)
- 34.123 K. Sośnica, D. Thaller, R. Dach, P. Steigenberger, G. Beutler, D. Arnold, A. Jäggi: Satellite laser ranging to GPS and GLONASS, *J. Geod.* **89**(7), 725–743 (2015)
- 34.124 J. Kouba, T. Springer: New IGS station and satellite clock combination, *GPS Solut.* **4**(4), 31–36 (2001)
- 34.125 F.J. Gonzalez Martinez: Performance of New GNSS Satellite Clocks, Ph.D. Thesis (Karlsruher Institut für Technologie, Karlsruhe 2014)
- 34.126 K. Senior: Report of the IGS working group on clock products, 19th Meet. Consult. Comm. Time Freq., Sèvres (BIPM, Sèvres 2012) pp. 219–236
- 34.127 J. Ray: REMINDER: Switch to IGS08/igs08.atx on 17 April 2011 IGSMail-6384 (2011) <https://igsb.jpl.nasa.gov/pipermail/igsmail/2011/007574.html>
- 34.128 J.R. Ray, J. Griffiths: Status of IGS orbit modeling and areas for improvement, *Geophys. Res. Abstr.* **13** (EGU, Vienna 2011) EGU2011-3774
- 34.129 S. Hilla: The Extended Standard Product 3 Orbit Format (SP3-c) (2010) <https://igsb.jpl.nasa.gov/igsb/data/format/sp3c.txt>
- 34.130 J. Ray, W. Gurtner: RINEX Extensions to Handle Clock Information (2006) https://igsb.jpl.nasa.gov/igsb/data/format/rinex_clock300.txt
- 34.131 J. Kouba, Y. Mireault: New IGS ERP Format (version 2), IGSMail-1943 (1998) <https://igsb.jpl.nasa.gov/mail/igsmail/1998/msg00170.html>
- 34.132 M. Rothacher and R. Schmid: ANTEX: The Antenna Exchange Format, Version 1.4 (2010) <https://igsb.jpl.nasa.gov/igsb/station/general/antex14.txt>
- 34.133 G. Gendt: IGS switch to absolute antenna model and ITRF2005, IGSMail-5438 (2006) <https://igsb.jpl.nasa.gov/pipermail/igsmail/2006/005509.html>
- 34.134 G. Maral, M. Bousquet: *Satellite Communications Systems: Systems, Techniques, and Technology*, 5th edn. (Wiley, Chichester 2009)
- 34.135 Radio Technical Commission for Maritime Services (RTCM): Differential GNSS (Global Navigation Satellite Systems) Services – Version 3 (2013)
- 34.136 S. Hackel, P. Steigenberger, U. Hugentobler, M. Uhlemann, O. Montenbruck: Galileo orbit determination using combined GNSS and SLR observations, *GPS Solut.* **19**(1), 15–25 (2015)

Surveying

35. Surveying

Chris Rizos

The Global Positioning System (GPS) became available as a civilian geodetic survey technology in the early 1980s. It has since revolutionized not only geodesy, but surveying operations as well. Global Navigation Systems (GNSSs) are today a fundamental tool for the land, engineering, and hydrographic surveyor. The majority of GNSS survey tasks relate to the determination of high-accuracy coordinates in a well-defined reference frame, typically using differential GNSS positioning techniques based on the analysis of carrier-phase measurements. Carrier-phase-based positioning is capable of distinct *levels* of accuracy – submeter, few decimeters, centimeter, and even subcentimeter – through a combination of special instrumentation, sophisticated software, and unique field operations. The evolution of GNSS from a geodetic surveying technology to a versatile surveying tool has seen precise positioning implemented in real-time, using ever shorter spans of measurements, and even when the user receiver is in motion. Furthermore, new techniques based on precise single-point positioning, as well as wide-area reference receiver networks, are starting to find wider use.

35.1	Precise Positioning Techniques	1013
35.1.1	Static Positioning	1014
35.1.2	Rapid-Static Positioning	1016
35.1.3	Kinematic Positioning	1017
35.1.4	Real-Time Differential GNSS Positioning	1019
35.1.5	Precise Point Positioning	1021
35.2	Geodetic and Land Surveying	1023
35.2.1	Geodetic Survey Applications	1023
35.2.2	Land Surveying Operations	1024
35.2.3	Land Surveying and Mapping Applications	1027
35.3	Engineering Surveying	1029
35.3.1	Engineering Surveying Real-Time Operations	1029
35.3.2	Engineering Surveying Applications	1030
35.3.3	Project Execution and Related Issues	1032
35.4	Hydrographic Surveying	1033
35.4.1	Hydrographic Surveying Applications	1033
35.4.2	Operational Issues	1035
	References	1035

Among the first civilian GPS user communities in the early 1980s were geodetic surveyors, who used the technology to determine the coordinates of ground marks in control networks. Today, around the world, GNSS is unchallenged as the primary technology for geodetic surveying.

Geodetic surveying requires the determination of geodetic coordinate information that is of high accuracy. This implies a level of coordinate accuracy significantly higher than that possible using standard GNSS open services, such as GPS's Standard Positioning Service (Chap. 7) or Galileo's Open Service (Chap. 9), which deliver meter-to-dekameter level single-point positioning accuracy (Chap. 21). In this chapter, the accuracy requirements for surveying and mapping applications will be assumed to be in the range from

subcentimeter to the submeter. Such high positioning accuracy requirements have spurred the development of unique observation procedures, measurement technologies, and data analysis methods – all of which are hallmarks of *GNSS surveying*.

High-accuracy GNSS positioning is synonymous with the differential positioning mode [35.1]. The differential GNSS techniques (Sects. 21.5 and 26.1) range from those based on pseudorange measurements to carrier-phase-based positioning which – depending upon the algorithm and operational mode that is used – can deliver accuracies from a few millimeters to several decimeters. New developments in precise point positioning (PPP; Chap. 25) offer an alternate mode of survey receiver operation that does not require a nearby simultaneously operating GNSS reference receiver.

One of the key features of differential GNSS techniques compared to terrestrial geodetic surveying techniques is that intervisibility between pairs of observing GNSS receivers is not necessary. In fact, the distance between GNSS receivers may range from a few kilometers for land or engineering survey applications, to hundreds and even thousands of kilometers in the case of global geodesy applications. Furthermore, the ground marks whose coordinates are to be determined are *static*. In the case of GNSS geodetic surveying great care is taken to build stable monuments upon which the GNSS antennas are mounted – concrete pillars, steel pins, metal tripods, or poles fixed to bedrock or attached to structures. The assumption is that the three-dimensional coordinates are determined once, and then these coordinated ground marks serve as the datum control marks to which all other (lower accuracy) surveys are *connected*. In this way, the datum or reference coordinate system is propagated to all geospatial data observed using any of the standard terrestrial or GNSS-based surveying and mapping techniques.

For many tasks, the geodetic, land, engineering, or hydrographic surveyor does not require coordinate information in *real-time* (RT). GNSS surveys typically have as their *raison d'être* the production of a digital map, the computation of the precise coordinates of the GNSS receiver antenna trajectory, or the establishment of a network of coordinated ground control marks. Nevertheless there are GNSS surveying applications where real-time coordinates are required, as in the case of machine automation applications, or for construction set-out tasks, or trajectory determination, or to navigate from one point to another (Chaps. 21 and 30).

It must be emphasized that *GNSS surveying* is actually an extension of *GPS surveying* – a set of precise satellite-based positioning techniques that have evolved over a period of about three decades [35.1–4]. In fact all mathematical concepts, measurement principles, operational procedures, and applications were first developed using GPS technology. With a heritage of geodetic survey applications, the first decade of GPS surveying was characterized by static positioning in which two GPS receivers recorded measurements during an *observation session*, and subsequent data processing generated the baseline vector connecting a ground point of known geodetic coordinate to a point whose coordinate was to be determined [35.2]. Back in the office, the recorded measurements from the pair of simultaneously operating receivers would be processed, one observation session at a time, to compute the single-session baseline vectors. A network of coordinated points observed in this way would be an effective realization of the geodetic datum, which could be used for subsequent survey and mapping tasks.

During the 1990s, a series of developments led to an extraordinary increase in the productivity of GPS:

- GPS surveying was enhanced by developments that offered an increased flexibility due to the short baseline survey mode.
- Rapid GPS positioning techniques, including real-time operations.
- Use of permanent GPS receivers (obviating the need for the surveyor to operate their own reference station receiver).
- High-accuracy (geocentric) geodetic national and regional datums.
- The availability of GPS data products such as those of the *International GNSS Service* (IGS; Chap. 33).

The drive for improvements in the performance of GNSS surveying techniques continues to this day. Key achievements include faster carrier-phase ambiguity resolution (AR; hence shorter observation sessions), more robust positioning (hence fewer erroneous baseline solutions), and lower operational constraints (hence lower field survey costs). These improvements result from several independent developments, such as multi-constellation GNSS (more satellites), more frequencies (more reliable AR, longer baselines), better designed signals (lower multipath), higher quality satellite clock and orbit data products, standardization of data file and transmission formats, permanent reference receiver networks, real-time carrier-phase-based techniques, geoid models (for height determination), and improved GNSS receiver technology.

Such improvements are not only of benefit to the geodetic and surveying community, but also they are facilitating the adoption of carrier-phase-based GNSS techniques into application areas such as machine guidance and automation (including robotics), rapid mapping (using terrestrial, marine and airborne sensors), construction and mining engineering operations, and precise navigation, to name but a few.

In summary, different GNSS positioning modes and data-processing strategies are all designed to account for systematic errors in the GNSS measurements, or contribute supplementary information for observation models, so as to assure a certain level of coordinate accuracy, at the minimum cost and complexity. The following have fundamental influences on the methods of GNSS positioning (Chaps. 21 and 26 and [35.1, 3, 4]):

1. The type of GNSS measurements – *carrier-phase* measurements are used because of their low noise.
2. Whether positioning is determined in an *absolute* sense using only *single-receiver* measurements, or

defining the position of one receiver *relative* to one or more reference receiver – the former implying the coordinate datum is fixed by satellite orbit information (as in the case of single point positioning or PPP); and the latter by the fixed/known coordinates of the reference receiver(s).

3. Whether the coordinated point is *stationary*, or is in *motion* – the former allows for a *stacking* of measurements that increase the solution redundancy (Chap. 22), and hence improve the precision (and, in general, the accuracy) of the estimated parameters; whereas the quality of kinematic positioning is strongly influenced by instantaneous satellite geometry and the magnitude of residual measurement biases or disturbances.
4. Whether the coordinate solution must be generated in *real-time*, or is derived *post-survey* – the former requires more complex instrumentation and additional infrastructure (variety of communication links, generation of real-time augmentation information, data formats, and protocols); whereas coor-

dinate solutions generated in post-survey mode are typically more accurate than those derived in real-time.

This chapter focuses on the precise positioning applications for geodetic, land, engineering, and hydrographic surveying, and is organized as follows. Section 35.1 introduces the fundamental classes of precise positioning techniques used for the various surveying applications, and discusses the characteristics of static and kinematic type positioning, as implemented in real-time or post-processing methods, based on either the relative or point positioning modes. Section 35.2 discusses the first of the GPS applications that used carrier-phase-based relative positioning techniques – geodetic surveying. All other forms of GNSS surveying have been derived from the basic geodetic surveying principles. Land surveying is introduced in Sect. 35.2. Sections 35.3 and 35.4 deal with engineering surveying and hydrographic surveying applications, respectively.

35.1 Precise Positioning Techniques

Civilian users have from the earliest days of GPS availability demanded ever increasing levels of performance, in particular higher accuracy, improved reliability, lower costs, and faster results. This is particularly true of geodesists, surveyors, and engineers, who seek accuracy that is several orders of magnitude higher than that required by other GNSS users. Although it is possible to categorize positioning applications according to a range of criteria, the following considerations are useful: accuracy, time sensitivity of positioning, time-to-coordinate-solution, receiver kinematics, infrastructure requirements, and nature of supplementary model information. Each of these is discussed below.

Accuracy traditionally has been expressed in relative terms, for example, as a ratio of coordinate error (typically expressed as a 95% uncertainty) to distance (between ground marks, or between GNSS receivers when operated in differential mode). The coordinate error then can be expressed in metric or distance units by scaling the ratio by receiver or ground mark separation. Hence *one part-per-million* (or 1 ppm) is a relative accuracy measure of one centimeter between two points separated by 10 km, or 0.5 cm over 5 km, or 10 cm over 100 km, etc. Furthermore, it can refer to a single coordinate component (e.g., *x*, *y*, or *z* Cartesian coordinates, or the *height component*) or a transformed coordinate quantity such as the *horizontal component*.

Surveys (and hence coordinates derived from them) were (and still are to a major extent) categorized in a *hierarchical* sense, from the highest *geodetic* categories through to lower accuracy control, engineering, and mapping surveys. Nowadays, the range of accuracies for high-accuracy GNSS surveys would be from subcentimeter to perhaps the decimeter-level. There is a complex relationship between, on the one hand, accuracy sought, and on the other hand the GNSS hardware, field procedures, and data-processing strategies that should be used. Some are formulated as recommended standards and guidelines; however many are not. Interestingly, the GNSS hardware varies the least, as invariably multi-frequency GNSS equipment is used no matter what type of survey is conducted (although there are different receiver/antenna form factors). In contrast, the measurement modeling used within the data-processing software varies considerably from commercial systems designed to satisfy the needs of land and engineering surveyors, optimized for rapid and easy use in constrained conditions (primarily with regards to length of observation time and inter-receiver distance), and geodetic software capable of ultra-high-accuracy intended for crustal motion and geoscientific applications (Chaps. 36 and 37).

Timeliness is a critical concern for some engineering and machine guidance applications, where the coordinate results are required without delay. This

gives rise to one of the most important distinguishing characteristics of high-accuracy GNSS: real-time operations or post-survey processing. The former has a considerable impact on operations and supporting infrastructure, whereas the latter is sufficient for geodetic applications, land and surveying, and most mapping needs. *Time-to-solution* is closely related to timeliness. Static geodetic survey operations typically require lengthy *observation sessions*; whereas high productivity and RT surveys must have very short *initialization* periods that subsequently enable precise single-epoch positioning. Long observation sessions are necessary for high-accuracy surveys over extended inter-receiver distances (hundreds to thousands of kilometers) typical of geodetic surveying applications. Hence time-to-solution is also closely related to the competitiveness of GNSS with conventional terrestrial surveying technology operating over typical survey project distances of the order of a few tens of kilometers or less.

Kinematics refers to the movement of the GNSS receiver while conducting the positioning task. A GNSS receiver may be in continuous motion, mounted on a variety of land, marine, air, and spaceborne platforms; or attached to a monumented ground mark; or perhaps in hybrid static–kinematic mode. The kinematic survey mode implies single-epoch, single-receiver positioning for each space point on a trajectory. On the other hand, static positioning (especially in post-processed mode) benefits from a massive increase in redundancy in positioning models [35.2] because many measurements can be used to determine the coordinates of a single stationary ground point. However, high-accuracy kinematic positioning capability is critical for many engineering surveying applications (Sect. 35.3).

High-productivity techniques for rapid surveying and real-time operations (e.g., in support of machine guidance, engineering, and construction) are very demanding of *reference receiver infrastructure*, including information technology and wireless communications. Relative positioning requires the operation of one or more *nearby* simultaneously operating reference receivers; whereas techniques such as PPP do not, in general, have this requirement. Furthermore, the density of *reference receiver networks* may vary from very low in the case of the most sophisticated geodetic static techniques or PPP to very high density (typically less than a few tens of kilometers spacing) for high productivity surveys and real-time operations.

Augmentation information is required by all precise positioning techniques, ranging from GNSS measurements at reference receivers in the case of relative positioning techniques to precise orbits, clocks, and perhaps atmospheric/bias information for PPP tech-

niques. With respect to supplementary model information, the critical distinction is between the transmission of augmentation information to users (with all the demands that places on infrastructure operations and service providers) and the provision of such information post-survey impacting upon the timeliness of precise positioning.

In the following sections, the major precise positioning techniques and their distinguishing characteristics are discussed in further detail. It must be emphasized that the development of multi-constellation GNSS receivers, and associated data-processing software, to take advantage of the massive increase in the number of available GNSS signals over the coming years (Chaps. 7–11), will lead to significant improvements in performance – from a reduction in observation session lengths for rapid-static surveys, to single-epoch AR, to relaxed (i.e., longer) user-reference receiver distance specifications (and hence lower infrastructure requirements), to increased reliability and quality of positioning. Furthermore, the increase in variety, access, accuracy, and applicability, and the decrease in latency of GNSS services will also lower the constraints for precise GNSS positioning. However, whether the cost of top-of-the-line geodetic-grade GNSS receivers will fall substantially is uncertain.

35.1.1 Static Positioning

With a heritage of geodetic survey applications, the first decade of GPS surveying was characterized by *static positioning*. The employed techniques can nowadays be generalized to static GNSS positioning and are summarized in Table 35.1. A survey with a minimum configuration of a pair of GNSS receivers progresses as follows [35.1]:

1. One (or more) receiver antenna would be set up on a monumented control point with known datum coordinates (the so-called *reference station* or *base station*), the other(s) over ground mark(s) whose coordinates are to be determined.
2. During an observation session, sufficient measurements of carrier-phase observations to the visible GNSS satellites would be recorded simultaneously by all receivers for a period ranging from an hour (or so) to several days.
3. One (or more) receiver would then be moved to another point and the antenna set up over a new ground mark. The other (or several) reference receiver(s) would occupy the same (or a new) datum control mark(s), and another observation session would ensure that measurements were recorded by the simultaneously operating receivers.

Table 35.1 Summary of precise GNSS positioning techniques – static positioning

- Typical scenario: two or more receivers used simultaneously in campaign, multisession mode, to record measurement files at many points
- Reference receiver separations are project-specific, ranging from tens to hundreds, and even thousands kilometers
- Observation session lengths from about an hour to several days; or continuous observations in case of permanent control or deformation monitoring points
- Monumentation: from highly stable to temporary ground marks
- Top-of-the-line GNSS receivers (carrier-phase measurements on at least two frequencies to form ionosphere-free observables), choke-ring (or equivalent) antennas
- Application typically for the establishment of geodetic control points, or densification of existing control marks
- Commercial software processing is in single-baseline mode, with simplified functional model of estimable parameters consisting of baseline (vector) components and double-differenced ambiguities (unresolved); requiring subsequent single-network adjustment of multiple baseline vectors
- Scientific (geodetic) software has rigorous multi-receiver, multi-session analysis capability; with options to estimate a wide variety of additional orbit, clock, bias, atmospheric, and reference frame parameters
- Resolution of ambiguities is not generally necessary
- Post-processed baselines or multi-receiver scenarios; with datum constrained by reference receiver coordinates; quality is a complex function of many environmental and observation factors, and the degree of sophistication of observation and reference frame modeling

Table 35.2 Comparison of static GNSS positioning techniques

	Single-baseline static GNSS surveying	Multi-station GNSS geodetic surveying
Datum	<ul style="list-style-type: none"> ● Base station per baseline processing ● Datum station(s) in network adjustment of baselines 	<ul style="list-style-type: none"> ● Small number of reference stations ● Typically IGS stations; International Terrestrial Reference Frame (ITRF)
Inter-receiver distances	<ul style="list-style-type: none"> ● Tens of kilometers 	<ul style="list-style-type: none"> ● 100–1000s km
Observation session	<ul style="list-style-type: none"> ● One to several hours 	<ul style="list-style-type: none"> ● Several hours to several days
Accuracy	<ul style="list-style-type: none"> ● Relative accuracy of 0.5–1 ppm horizontal, 1–2 ppm vertical; implying centimeter-level coordinate accuracy over typical baseline lengths 	<ul style="list-style-type: none"> ● 1–10 ppb; implying centimeter-level coordinate accuracy within GNSS networks over 100–1000s km extents
GNSS hardware	<ul style="list-style-type: none"> ● Single-frequency GPS; or multi-GNSS, multi-frequency receiver ● Light-weight antenna, mounted on tripod 	<ul style="list-style-type: none"> ● Multi-GNSS, multifrequency receiver ● Choke-ring (or equivalent) antenna, mounted on stable monumentation
Processing	<ul style="list-style-type: none"> ● Commercial off-the-shelf baseline processing software; automatic processing ● Receiver INdependent EXchange (format) (RINEX) or proprietary data files ● Simplified functional model 	<ul style="list-style-type: none"> ● Multi-receiver, multi-station scientific software; considerable analyst skill ● Web processing (automatic) ● RINEX data and auxiliary model or information files ● Sophisticated functional model
Estimated parameters	<ul style="list-style-type: none"> ● Baseline vector ● Double-differenced, unresolved (real-valued) ambiguities ● Following network adjustment: individual receiver coordinates 	<ul style="list-style-type: none"> ● Receiver coordinates ● Ambiguities, tropospheric parameters ● Optionally satellite orbits, biases, Earth orientation parameters, etc.
Applications	<ul style="list-style-type: none"> ● Project control surveys; and other postprocessed surveys ● Alternative to terrestrial control survey technologies 	<ul style="list-style-type: none"> ● Reference frame observations ● Geodynamics and other geodetic applications

4. This procedure of moving receivers to predefined points, and recording measurements made at all GNSS receivers, would be repeated until all ground marks in the survey area were visited at least once – always ensuring that there was a *link*, or baseline connection, back to one or more datum control points.

There are two classes of static GNSS relative positioning techniques, which are compared in Table 35.2.

On the one hand there are the ultra-accurate, long baseline GNSS techniques – capable of relative positioning accuracies of tenths of ppm up to several parts-per-billion (ppb) over baseline lengths of hundreds to thousands of kilometers. The measurements are made by top-of-the-line, multi-frequency, multi-constellation GNSS receivers and the observation sessions last for many hours or even days. The measurement processing is undertaken using sophisticated scientific software executed in post-survey mode to support a series of global or national geodesy applications (Chaps. 36–39). As an alternative to processing the measurement data themselves – a task that requires considerable analyst skill – surveyors can submit observation data files in the Receiver Independent Exchange (RINEX) format (Annex A.1.2; [35.5]) to one of several web processing engines such as NGS's OPUS [35.6], NRCAN's CRCS-PPP [35.7], GA's AUSPOS [35.8], and others.

At the other end of the spectrum are the medium-to-short baseline GNSS survey techniques. They are capable of accuracies of a few ppm for baselines perhaps up to several tens of kilometers in length and are typically employed to support control network applications. Although low-cost single-frequency hardware could, in principle, be used, today's GNSS surveying hardware is essentially the same *geodetic-grade* receivers as would be used for any of the multi-frequency precise positioning techniques [35.2]. However, the measurement processing is carried out using commercial software packages provided by GNSS receiver manufacturers, distinguished from the scientific software referred to earlier by the use of significantly simplified GNSS observation modeling. In such scenarios the recorded measurements from a pair of simultaneously operating receivers would be processed, one observation session at a time, to compute single-session *baseline* vectors. Following baseline processing, carried out for baselines during each independent observation session, the multiple computed baselines would undergo a *secondary network adjustment* [35.3, 4]. The three-dimensional (3-D) baseline vectors are in effect treated as the observations to be adjusted – with the output being the optimal coordinates of the entire ground control network constrained by datum

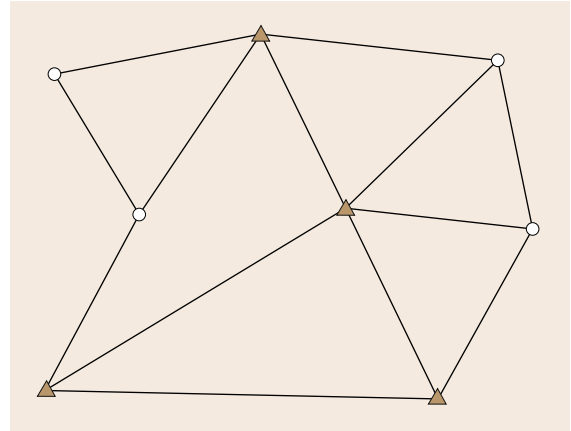


Fig. 35.1 From independent GNSS baselines to survey network: a network of coordinated points is constructed by linking together separate baselines (i.e., pairs of simultaneously operating GNSS receivers) that connect and propagate the known coordinates of ground control marks (*triangles*) to other points whose coordinates are to be determined (*circles*). Each baseline links a GNSS receiver at a point of known coordinates (available a priori or estimated from a GNSS baseline solution) and a receiver whose coordinates are to be determined. Extra baselines can provide redundant pathways of generating coordinate information for quality control purposes

control points. Such a network of coordinated points can be used for subsequent survey and mapping tasks (Fig. 35.1).

Conventional static GNSS positioning techniques are characterized by long observation sessions. Although an effective means of mitigating residual systematic biases, multipath, and model errors, this imposes significant constraints for routine surveying applications. Over the last two decades several precise GNSS surveying techniques and methodologies have been developed with the following *liberating* characteristics: (a) static antenna setups not required, (b) long observation sessions not essential, and (c) coordinates could be determined in the field. Each is a technological solution to the challenge of ensuring high productivity (coordinating as many points in as short a field survey time as possible) and/or versatility (e.g., the ability to obtain results even while the receiver is in motion and/or in real-time) without sacrificing coordinate accuracy and solution reliability.

35.1.2 Rapid-Static Positioning

For *rapid-static* positioning (Table 35.3) observation session lengths are significantly shorter than for conventional static GNSS surveying discussed above. Ob-

Table 35.3 Summary of precise GNSS positioning techniques – rapid-static positioning

- Typical scenario: one user receiver in single-baseline configuration
- Reference receiver may be operated by user or by third party, on a single project basis or as continually operated reference receiver; datum defined by reference receiver coordinates
- User-reference receiver separations typically tens of kilometers, often < 10 km for very fast surveys
- Observation session lengths from a few minutes to over 30 mins, but must be sufficient for ambiguities to be resolved
- Monumentation standards are project specific
- Multi-frequency GNSS receivers (carrier-phase and pseudorange measurements), light-weight (portable) antennas
- Applications are typically the coordination of many ground marks, minor control, detail, or as-built surveys
- Data processing via commercial software; may also be undertaken in real-time mode
- Relatively fast high-accuracy GNSS surveying tool
- Assuming use of multi-frequency receiver, quality of solution is a function of baseline length (degree of cancellation of spatially correlated biases), observation session length, quality of measurements, number of tracked satellites, sophistication of data-processing algorithm

servation session length is a complex function of user-reference receiver baseline length, number of multi-frequency measurements, number of satellites tracked, satellite geometry, and presence of multipath disturbances. Accordingly, hard and fast rules are impossible to formulate. Typically, however, receivers need only to occupy a station for a period of perhaps a few minutes for baselines of less than 10 km in length and good satellite coverage. Here, *good* refers to the overall number of tracked satellites (a minimum of six is generally sufficient) and their distribution across the sky (i. e., satellites should be observed in at least three of the four NE-SE-SW-NW quadrants). Extended observation sessions of perhaps up to 15 min or more may be required for longer baselines, less tracked satellites, and/or poor sky distribution of satellites. Several references to GNSS survey guidelines with recommendations regarding observation session length are provided in Sect. 35.2.2. When utilizing measurements from the full complement of GNSS constellations, on two or more signal frequencies, it is expected that the length of the observation session will reduce dramatically, perhaps even down to a single-epoch.

The basis of the rapid-static positioning technique is the ability of the measurement-processing software to resolve the ambiguities using a *very short observation session* – the data analysis software must therefore have a *rapid AR* capability (Chap. 23). The rapid-static field procedures are similar to those for conventional static GNSS surveying, except that: (a) observation session lengths are *shorter*, (b) the baselines are comparatively *short*, (c) the satellite geometry needs to be *favorable*, and (d) signal disturbances such as multipath should be *minimal*. While the observation of independent baselines is the same survey scenario as in the case of conventional static GNSS positioning using commercial software, another more common scenario is the determination of *radiations* of vectors from a sin-

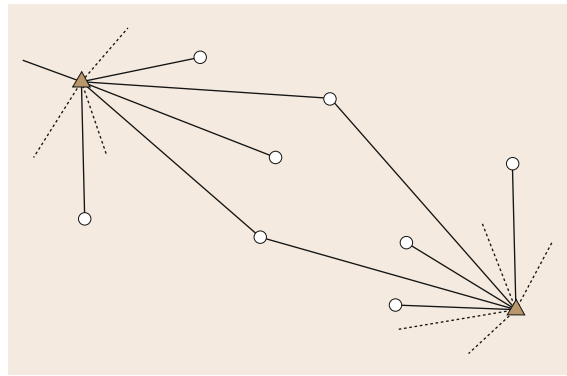


Fig. 35.2 Geometry of rapid-static baselines: static ground points (*circles*) can be coordinated by the differential GNSS positioning mode via the measurement of 3-D baselines connected to base stations with known coordinates (*triangles*). One can see how logistically efficient this method of observing *radiating* baselines is when a GNSS receiver only needs to occupy the ground marks for short measurement sessions. Note the use of two base stations increases opportunities for quality control. Multiple base stations may or may not be operated simultaneously

gle (or two or three) reference stations as indicated in Fig. 35.2.

The rapid-static technique is well suited for short-range applications such as establishing project-scale control and for certain types of land surveys (Sect. 35.2.3). The essential characteristics of rapid-static GNSS positioning are summarized in Table 35.4.

35.1.3 Kinematic Positioning

Table 35.5 lists some of the characteristics of *kinematic GNSS positioning*. We may distinguish between two forms of kinematic positioning. The first is when the coordinates of the moving GNSS receiver antenna's

Table 35.4 Characteristics of rapid-static and conventional static GNSS positioning

	Rapid-static GNSS surveying	Conventional static GNSS surveying
Datum	<ul style="list-style-type: none"> ● Single base station ● Radiation of baseline vectors from single reference receiver 	<ul style="list-style-type: none"> ● Base station per baseline processing ● Datum station(s) in network adjustment of baselines
Inter-receiver distances	<ul style="list-style-type: none"> ● Typically less than conventional static 	<ul style="list-style-type: none"> ● Tens of km
Observation session	<ul style="list-style-type: none"> ● Few minutes to < 1 h; see <i>factors impacting on accuracy</i> 	<ul style="list-style-type: none"> ● One to several hours
Accuracy	<ul style="list-style-type: none"> ● 1–2 cm horizontal, 2–3 cm vertical; over typical baseline lengths 	<ul style="list-style-type: none"> ● 0.5–1 parts-per-million (ppm) horizontal, 1–2 ppm vertical; i. e., centimeter-level accuracy over typical baseline lengths
GNSS hardware	<ul style="list-style-type: none"> ● Multi-GNSS, multi-frequency receiver (preferred) ● Light-weight antenna, mounted on tripod 	<ul style="list-style-type: none"> ● Single-frequency GPS (lower performance); or multi-frequency receiver (preferred) ● Light-weight antenna, mounted on tripod
GNSS software	<ul style="list-style-type: none"> ● Commercial off-the-shelf baseline processing software; automatic processing ● RINEX or proprietary data files ● Simplified functional model; <i>rapid AR capability</i> 	<ul style="list-style-type: none"> ● Commercial off-the-shelf baseline processing software; automatic processing ● RINEX or proprietary data files ● Simplified functional model
Estimated parameters	<ul style="list-style-type: none"> ● Baseline vector ● Resolved ambiguities, i. e., ambiguity-fixed baseline solutions ● Quality control implemented via re-visit of ground marks 	<ul style="list-style-type: none"> ● Baseline vector ● Double-differenced, real-valued ambiguities ● Following network adjustment: individual receiver coordinates
Factors impacting on accuracy	<ul style="list-style-type: none"> ● Baseline length ● Observation session length ● Quality of carrier-phase and pseudorange measurements ● Multi-frequency measurements ● Number of tracked satellites and geometry 	<ul style="list-style-type: none"> ● Baseline length ● Observation session length ● Quality of carrier-phase measurements
Applications	<ul style="list-style-type: none"> ● Project control surveys ● Detail, as-built, and other post-processed surveys ● Alternative to terrestrial survey technology 	<ul style="list-style-type: none"> ● Project control surveys; and other post-processed surveys ● Alternative to terrestrial control survey technology

Table 35.5 Summary of precise GNSS positioning techniques – kinematic positioning

- Typical scenario: one mobile user receiver in single-baseline configuration; on a variety of land, marine, aerial, or spaceborne platforms
- Reference receiver may be operated by user or by third party, on a single project basis or as continually operated reference receiver; datum defined by reference receiver coordinates
- As with rapid-static surveys, user-reference receiver separations typically tens of kilometers, often < 10 km to ensure AR using geodetic-grade GNSS receivers with short observation sessions
- Single-epoch positioning using double-differenced *carrier-range* observable (double-differenced carrier phase with integer resolved ambiguities), also known as ambiguity-fixed solutions
- Options for re-initialization must be available if signal loss-of-lock on five or more satellites, and may include remaining stationary until ambiguities resolved again, return to previously surveyed static point, etc.
- Multi-frequency GNSS receivers, light-weight (portable) antennas
- Applications typically are the coordination of receiver antenna trajectories, for example, for mapping projects (road centerline surveys, aerial imaging/scanning), satellite orbit determination, hydrographic charting, etc.
- Model and data processing via commercial software; may also be undertaken in real-time mode (see below)
- Assuming multi-frequency user receiver, quality of solution is a function of baseline length (magnitude of residual differential biases), correctness of AR process, quality of measurements (i. e., multipath-free), number of tracked satellites, satellite-receiver geometry (i. e., Dilution of Precision measures)



Fig. 35.3 Precise kinematic GNSS survey: the receiver is installed on a quad-bike with the antenna mounted upon on a pole, and the kinematic positioning task consists of determining coordinates of the antenna on a continuously sampled basis (e.g., once per second) as the bike is driven up and down the beach, to determine a dense network of heights with centimeter-level accuracy for beach erosion studies; note that the antenna height must be corrected for the fixed height of the top of the pole above the ground level (courtesy of Brad Morris)

trajectory is required (as in Fig. 35.3). The second category is a form of static positioning, but with the special case that the receiver continues to track satellites while it is moved from one static point to another.

The *stop-&-go* GNSS surveying technique deserves special consideration because the coordinates of the receiver are only of interest when it is stationary (the *stop* part); however the receiver continues to function while it is being moved (the *go* part) from one stationary setup to the next, as is indicated in Fig. 35.4.

The first step that needs to be performed in the survey is the initial AR in order that all subsequent single-epoch solutions are based on carrier-range positioning (Sects. 23.2 and 26.3). This technique is well suited to projects where many points close together have to be surveyed, and the terrain does not cause significant signal obstructions.

Instead of only coordinating the stationary points and disregarding the trajectory of the roving antenna, the objective of *kinematic* surveying is to determine the position of the antenna while it is in motion. In many other respects the technique is similar to the *stop-&-go* positioning technique. That is, the ambiguities must be resolved *before* starting the survey, and the ambiguities must be re-initialized *during* the survey when loss-of-signal-lock occurs which causes the ambiguity parameters to change from their initial values. Kinematic positioning invariably involves the determination of vectors radiating from a single (or small number

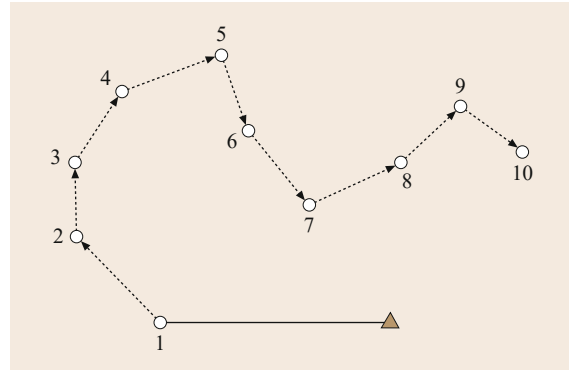


Fig. 35.4 Progress of a *stop-&-go* GNSS survey: the first baseline is observed (known control mark to point 1), and once the ambiguities have been resolved (e.g., using the rapid-static positioning technique), the user receiver's antenna is then moved carefully from point 1 to point 2, then to point 3, and so on, making just a few seconds of measurements while stationary at the ground point (circle). Note the base station (triangle) operates continuously and the point coordinates are determined by the radiated baseline method (Fig. 35.2); the trajectory of the antenna is not of interest, only the coordinates of the stationary points 1–2–3–4...

of) base or reference station(s) (Fig. 35.2). Kinematic GNSS surveying techniques are appropriate for road centerline, topographic and hydrographic surveys, airborne applications, etc.

35.1.4 Real-Time Differential GNSS Positioning

Real-time kinematic (RTK) GNSS is a popular technique for many survey applications as there is no post-processing of GNSS measurement data (Sect. 26.3). The standard differential positioning scenario as before requires the use of a pair of GNSS receivers connected by a wireless data link (Table 35.6). Successful operation of RTK-GNSS systems using radio modem data links is typically limited to baseline lengths of 5–10 km due to radio range constraints. Wireless links over the mobile Internet do not have such distance restrictions. However, the inter-receiver distance over which *rapid* AR algorithms work reliably using dual-frequency GNSS instrumentation (with good sky visibility) may only be 20–30 km, and often less in the event of high ionospheric activity (Chap. 39). As with carrier-phase-based kinematic positioning in general, when signals are obstructed the AR algorithm has to be restarted in order to resolve the (new) ambiguity parameters. As this may take several tens of seconds, and if signal interruptions occur frequently, then this

Table 35.6 Summary of precise GNSS positioning techniques – real-time differential positioning

●	Rapid-static and kinematic may be conducted in real-time, and collectively these ambiguity-fixed, short-baseline approaches are known as <i>real-time kinematic</i> (RTK) techniques
●	Operational constraints, reference receiver infrastructure requirements, and GNSS receiver specifications are as for rapid-static and kinematic GNSS surveys (see above)
●	May distinguish between operational constraints for <i>single-base RTK</i> (with baseline lengths of a few tens of kilometers), and multiple reference receiver <i>network-RTK</i> with sparser reference receiver network surrounding user receiver (50–100 km spacing) [35.9]
●	Additional infrastructure: networked reference receivers, analysis or network operations facility, and communications links (between reference receivers, and between reference receiver, or RTK service facility, and user receiver)
●	Variety of wireless communication links, though increasingly via mobile Internet (terrestrial or satellite) channels; with interoperability afforded by use of industry standard data transmission messages and protocols such as Radio Technical Commission for Maritime Services (RTCM)
●	Applications include all those that require precise coordinates in real time, to guide a machine or vehicle, for engineering or construction, and others
●	Versatile high-accuracy GNSS positioning technique when supported by the necessary augmentation infrastructure
●	Factors impacting quality are as for kinematic surveys, and in addition the reliability of the RTK communications link

dead time can result in RTK-GNSS being a comparatively inefficient positioning technique. The advantage over post-processed implementations of precise kinematic positioning is that when operated in real-time, the GNSS controller is able to alert the user in the event of the need for ambiguity re-initialization (i. e., new AR), or if there is an interruption in wireless communications from the reference receiver or RTK services center.

Most users subscribe to RTK-GNSS services rather than running their own reference receiver. Real-time networks of continuously operating reference stations (CORSs) have been established since the mid-1990s, and there are few signs of this trend slowing. One of the drivers for CORS investment, and the promotion of the use of RTK-GNSS, is the adoption of industry standard RTCM data message format and protocols (Annex A.1.3; [35.10]), ensuring interoperability between different brands of reference and user GNSS receivers. CORS installations typically comprise top-of-the-line receivers, with choke-rings antennas, capable of making multi-frequency, multi-GNSS measurements. The challenge is to install CORSs at sufficient density (minimum reference receiver separation) to permit single-base RTK with rapid AR (see schematic in Fig. 35.5). Note that this density of CORSs is the same as that for post-processed rapid-static and kinematic GNSS positioning using relative positioning principles.

RTK-GNSS implementations based on a network of reference stations (rather than a single reference station) are now common in many countries. Recall that one of the primary purposes of reference stations is to mitigate the impact on coordinate solutions of those systematic measurement biases and model errors that are spatially correlated [35.11, 12]. (The other is to

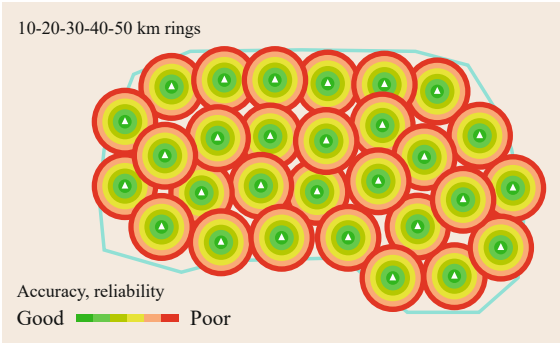


Fig. 35.5 CORS infrastructure for single-base RTK – indicating the *packing* of a network of continuously operating reference stations to support single-baseline, rapid-static, and kinematic positioning, so as to provide complete coverage over an area. Note the closer the user receiver is to a CORS the more reliable the GNSS position solution. Ideally the distance should be no greater than a few tens of kilometers (hence the graduation in color of the rings around each CORS indicating varying solution quality, from *green* for the highest to *red* for the lowest)

provide the datum for differential positioning.) In the simplest configuration, it is assumed that in the case of two *nearby* GNSS receivers, when the measurements are made at the same time, and processed in an integrated observable model such as that produced by double-differencing measurements from a pair of receivers to a pair of GNSS satellites, there is no (or negligibly small) effect of atmospheric refraction biases and satellite orbit/clock errors on the baseline results (Sects. 21.3 and 26.1). Of course that assumption breaks down as the distance between the two GNSS receiver increases. Hence single-base RTK essentially requires

Table 35.7 Comparing single-baseline RTK and network-RTK GNSS positioning

	Single-baseline RTK	Network-based RTK
CORS infrastructure	<ul style="list-style-type: none"> ● User owned CORSs; or Service Provider (SP) owns CORS ● Many CORSs for full area coverage 	<ul style="list-style-type: none"> ● Service Provider (SP) owns CORSs; or licenses raw data from organization operating network of CORSs ● Evenly distributed CORSs across service area
Service provision	<ul style="list-style-type: none"> ● User-operated; or SP ● RTCM v2 or v3 messages 	<ul style="list-style-type: none"> ● SP; or CORS operator ● RTCM v3 N-RTK messages; or VRS-based customized RTCM v2 messages
Inter-receiver distances	<ul style="list-style-type: none"> ● Tens of kilometers, preferably < 10 km for most reliable operations and/or rapid on-the-fly AR 	<ul style="list-style-type: none"> ● 50–100 km CORS spacing across service area ● < 30–50 km user receiver to nearest CORS
Configuration	<ul style="list-style-type: none"> ● Nearest CORS ● Owner-operated; or subscription to SP ● Typically direct user-CORS connection 	<ul style="list-style-type: none"> ● User located within cluster of 3–4 CORSs ● Subscription to SP ● Central network or operations server/facility
Modeling of spatially correlated biases	<ul style="list-style-type: none"> ● Cancellation of satellite-specific and atmospheric biases in double-differenced measurement model, by assuming biases identical to those of nearest CORS ● RTCM messages are <i>calibration</i> of biases at CORS location 	<ul style="list-style-type: none"> ● Cluster of CORSs surrounding user receiver location used to derive bias <i>correction surface</i> ● RTCM messages carry all information necessary for computation of location-specific biases to be applied as corrections to user receiver measurements
Communication options	<ul style="list-style-type: none"> ● Terrestrial: ultra-high frequency (UHF), very high frequency (VHF), MF beacons, digital broadcasts, mobile Internet, etc. 	<ul style="list-style-type: none"> ● Terrestrial: mobile Internet ● Satellite communications
Accuracy	<ul style="list-style-type: none"> ● Centimeter-level horizontal accuracy; 2× worse for vertical 	<ul style="list-style-type: none"> ● Similar accuracy to RTK; though with higher solution reliability
GNSS hardware	<ul style="list-style-type: none"> ● Single-frequency GPS (but distances shorter and/or longer AR process); or multi-frequency receiver (preferred) ● Light-weight antenna, mounted on bipod or pole, or moving platform 	<ul style="list-style-type: none"> ● Multi-GNSS, multi-frequency receiver (preferred) ● Same antenna and mounting options as RTK
Applications	<ul style="list-style-type: none"> ● Engineering surveying, machine guidance, and control (in agriculture, mining, construction, port operations, etc.) 	<ul style="list-style-type: none"> ● Same as RTK but operations less constrained by distance to nearest CORS; increased reliability due to multi-CORS configuration

the determination of baseline vectors radiating from the nearest RTK-capable reference station.

In contrast, a cluster of CORSs can be used to map the spatially correlated biases and errors across a CORS coverage area, and to apply these corrections to measurements at the user receiver location. This multi-CORS RTK-GNSS technique is often referred to as *network-RTK* (N-RTK), and its primary advantage from the point of view of the RT GNSS service providers is that the separation between user receivers and the surrounding CORSs can be of the order of 50–70 km or more [35.11–14]. The assumption that the systematic biases and model errors at the nearest CORS are the same as those at the user receiver location is replaced by the more realistic assumption that the biases and errors can be predicted at the user receiver location using a model of these biases and errors [35.15, 16].

N-RTK services can be supported by less dense CORS networks than single-base RTK services. Fur-

thermore, the precise coordinates are not strictly determined relative to a single (or nearest) reference receiver, but has similarities to network-based (i.e., multi-station) static positioning. There are a number of implementations of N-RTK [35.14, 16, 24–26], of which the virtual reference station (VRS) scheme is the oldest and best known [35.15]. Some characteristics of RT-GNSS positioning are summarized in Table 35.7.

As a result of substantially more measurements and frequency-diversity, multi-constellation precise GNSS positioning will be able to be carried out with significantly greater distance-to-CORS than current N-RTK implementations – distances of over 100 km.

35.1.5 Precise Point Positioning

PPP (Chap. 25 and [35.19, 27–29]) is a GNSS carrier-phase-based positioning technique that can be used anywhere on the globe by a single user receiver – at

Table 35.8 Summary of precise GNSS positioning techniques – precise point positioning

- Typical scenario: one static or mobile user receiver; the latter on a variety of land, marine, aerial or space platforms
- User GNSS receiver may be single-frequency or multi-frequency; the latter being preferred as such hardware is identical to survey-grade GNSS receivers used for differential carrier-phase-based positioning
- Requires precise satellite orbit and satellite clock information from an external source; post-processed orbit and clock information available in the form of several open standard formats such as Standard Product 3 (SP3; Annex A.2.1; [35.17]); RT streams use open RTCM-SSR messages (Annex A.1.3; [35.10]) or proprietary messages
- No reference receiver requirements for user positioning (although a sparse global network of reference receivers are needed for the computation of satellite orbits and clocks)
- Datum defined by reference frame in which orbits, clocks, biases and other parameters are computed; typically the International Terrestrial Reference Frame (ITRF) (Chaps. 2 and 36; [35.18])
- Real-time or post-processed software does not require reference receiver measurements; observation modeling is more sophisticated (and complete) as it must account for all systematic biases and model effects (those that may have been mitigated or eliminated in between-receiver data differencing)
- Dual-frequency PPP is capable of providing accurate position solutions at subdecimeter level for kinematic positioning and at subcentimeter level for static positioning [35.19, 20]. For single-frequency PPP the positioning accuracy lies at the decimeter level with high-end receivers [35.21] and for kinematic positioning with low-end receivers at the submeter level [35.22]. The high accuracy of dual-frequency PPP is achieved after a relatively long convergence time in the range of 20–40 min, while the decimeter level accuracy of single-frequency PPP is achieved in minutes [35.23]. Additional infrastructure, in the form of networked reference receivers similar to that for N-RTK operations, is necessary for rapid convergence or reliable AR
- Communication links for RT-PPP include geostationary satellite communications, mobile Internet, and downlink messages on navigation satellite signals
- Applications include all those that cannot be easily addressed using relative GNSS positioning techniques, including operations in remote and offshore areas
- Factors impacting quality are similar to those for kinematic surveys, such as satellite-receiver geometry, number of satellites tracked and measurements that are made, but also quality of the model information, algorithm and whether AR is successful (or even required)

least without direct co-processing of CORS measurements, or application of differential correction or model information generated from such measurements (Table 35.8). PPP offers, therefore, considerable flexibility, making it well suited for remote locations (on land and offshore) where there is an absence of GNSS CORS infrastructure.

PPP relies on accurate satellite orbit and clock error information that can be obtained from sources such as the International GNSS Service (Chap. 33), or a number of commercial service providers, and the explicit modeling of a number of measurement biases and system effects that are assumed to have been eliminated when using GNSS in relative positioning mode. PPP can be implemented in post-processed mode or in real-time. The former uses accurate orbit and clock products that are available – depending upon the product that is used – immediately or up to several weeks after the survey task is executed. As an alternative to doing their own processing, surveyors can submit RINEX data files to one of several web processing engines [35.6–8]. The latter uses RT orbit and clock data streams broadcast via the Internet (in the case of the IGS Real-Time Service – IGS-RTS [35.30, 31]), or satellite communications links. These streams may be in proprietary message formats

or in the RTCM State Space Representation (SSR) format (Annex A.1.3; [35.10]).

PPP can be done with single- and multi-frequency receivers. Fast single-frequency PPP requires next to orbits, clocks, and differential code biases, also ionospheric maps [35.23, 32–34]. The best possible position accuracy, a few centimeters or better, is obtained by using carrier-phase measurements from dual-frequency receivers. However, single-frequency receivers can provide decimeter accuracy at a reduced cost for the receiver and generally reach this level of accuracy much faster (few minutes) than a dual-frequency receiver does [35.23]. The convergence of dual-frequency PPP is longer (20–40 min) than that of single-frequency PPP as it initially depends on the relatively noisy ionosphere-free linear combination of the code data. Of course, after some time, the ionosphere-free linear combination of the carrier-phase data kicks in for dual-frequency PPP and becomes the determining factor for its high positioning accuracy.

The relatively long convergence time to reach the subdecimeter positioning accuracy, is one of the weaknesses of PPP. The positioning concept of PPP-RTK aims to address these weaknesses by reducing convergence times and improving positioning accuracy [35.35, 36]. It extends the PPP concept by pro-

viding single-receiver users, next to the orbits and clocks, also information about the satellite-phase biases. This information enables recovery of the integer user-ambiguities, thus enabling single-receiver AR thereby reducing the convergence times as compared to that of PPP. At present various different mecha-

nizations of PPP-RTK are under development [35.37–39]. When combined with atmospheric corrections, PPP-RTK is rivalling the speed of standard N-RTK (Table 35.7). This is a fertile area of GNSS research and substantial improvements in performance are expected.

35.2 Geodetic and Land Surveying

With the progressive refinement of GPS geodetic surveying techniques to make them easier to use, and to increase their versatility, it was inevitable that the application of GPS technology would extend to include land (see below), engineering (Sect. 35.3), and hydrographic surveying (Sect. 35.4).

35.2.1 Geodetic Survey Applications

Geodetic surveying was the first civilian application of precise GPS positioning [35.3, 4]. It is concerned with the establishment, maintenance, and densification of *geodetic datums* – across a range of scales from the global, regional, national, and state territory down to an individual project application (though these are sometimes referred to as *control surveys*, Sect. 35.2.3). Geodetic datums are realized by ground marks with known ellipsoidal coordinates that can be used by any surveyor or engineer as starting coordinates for subsequent precise surveys, to support mapping, surveying, construction, or engineering activities. Precise GPS static positioning (Sect. 35.1.1) revolutionized geodetic surveying because it was able to replace the traditional, slow, labor-intensive terrestrial surveying techniques.

There are several innovations of modern GNSS geodetic surveying methodology that bear mentioning. The first one is the near universal installation of permanent GNSS reference receivers, or CORSs. CORSs range from single-receiver installations to vast networks of CORSs across entire countries (as in Japan's GEONET [35.40], Sweden's SWEPOS [35.41]), regions (e.g., EUREF's permanent CORS network [35.42]), and globally (e.g., the IGS network [35.43]).

Figure 35.6 illustrates some of the CORS sites across the continent of Australia. Note that this is a non-homogeneous CORS networks, with different agencies and individuals being responsible for their operation. This is typical of many national CORS networks. In effect such networks consists of numerous *subnetworks* of CORSs, some established by the federal government agency responsible for geodesy, some by state government departments, and others by private companies,

local government authorities, universities, and even individual users. Furthermore the subnetworks may have different equipment configurations, different types of antenna mounts, and monumentation; supporting different user groups with a variety classes of service.

Figure 35.7 shows a typical choke-ring antenna (Chap. 17) with and without radome, installed on two typical designs of geodetic-grade monuments: concrete pillars and rigid tripods. The CORS coordinate reference point may not be the electrical center of the antenna but instead a physical reference mark on the top of the stable monument. Not shown is the instrument cabinet where the receiver itself is housed (together with communications, batteries, and other ancillary equipment), power systems such as solar panels, lightning protection, additional pillars or witness marks, etc. CORS installations such as these are a considerable in-

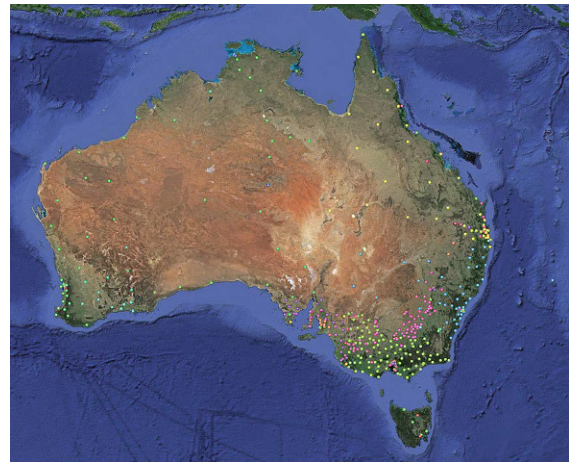


Fig. 35.6 GNSS CORS sites across Australia: an example of national CORS infrastructure that is non-homogeneous, with different owners and operators (*different colored dots*), to support a variety of GNSS positioning applications (and techniques), with an uneven distribution across the continent; note that private CORS sites operated by individual farmers, mining companies, universities, local council authorities, and others, are not shown (courtesy of Grant Hausler and ThinkSpatial)

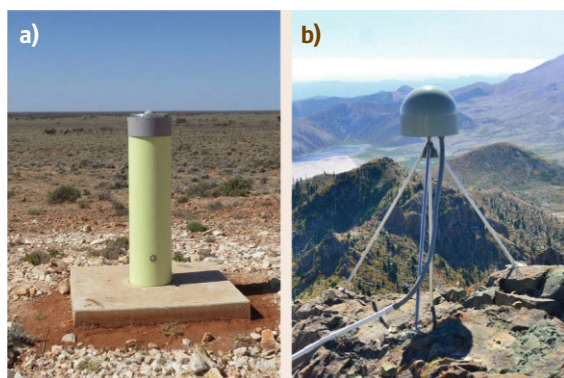


Fig. 35.7a,b Examples of geodetic-grade CORS installations: (a) on concrete pillar at Mulgathing, in South Australia, part of the AuScope national GNSS network (courtesy of Geoscience Australia); (b) drilled-braced monument at Coldwater Peak, part of the EarthScope Plate Boundary Observatory Mount St. Helens subnetwork (courtesy of Michael Gottlieb, UNAVCO)

vestment by an agency or organization in GNSS ground infrastructure.

The second innovation has been the availability of a variety of geodetic products and services, including those provided directly by the IGS (Chap. 33), by web services for GNSS measurement processing (Sect. 35.1.1), and by service providers for RT-GNSS positioning (Sect. 35.1.4), as well as the establishment of standardized data and transmission formats (Annex A) that support GNSS interoperability. This has also had an impact on GNSS measurement processing, allowing for the use of commercial software packages – as opposed to *scientific software* – for all but the most precise, long-baseline geodetic surveys.

The third concerns the nature of geodetic datums themselves. Increasingly national datums are aligned to, or defined by, the highest fidelity global geodetic datum: the International Terrestrial Reference Frame (ITRF; [35.18]). There are a number of reasons for this trend: (a) the global applicability of the ITRF, (b) the very precise set of coordinates and velocities of many GNSS CORSs (such as the IGS's network), (c) the ease of access via CORS tracking data and IGS geodetic products, (d) the well-defined datum epoch and documented maintenance procedures, and (e) its maintenance to the highest standards by the International Earth Rotation and Reference Systems Service (IERS; [35.44]).

Several GNSS geodetic surveying *methodologies* that can be used to distinguish these types of applications from routine engineering and mapping applications that rely on static and kinematic GNSS positioning techniques are summarized in Table 35.9. Note that

geodetic surveying assumes the use of *geodetic-grade* multi-frequency, multi-GNSS receivers, with choke-ring or multipath-mitigating antennas, set-up on stable monummentation (Fig. 35.7).

The geodetic survey applications may, at first glance, appear simply as examples of static surveys; however modern geodesy recognizes that no object on the surface of the Earth has zero velocity with respect to the ITRF. The mission of modern geodesy is to determine and monitor the coordinates of sample points in order to improve our knowledge of geophysical processes that have ground motion/deformation signatures [35.45, 46].

Ground deformation surveys are undertaken to measure the change in the coordinates of stable points or monuments fixed to the Earth's surface. The points may move in a horizontal or vertical sense, or in three dimensions, with signature characteristics across a wide range of time and spatial scales, from continental motion of the order of millimeters or centimeters per year, to rapid ground shaking during an earthquake reaching magnitudes of many decimeters. There are a number of subcategories of deformation surveys, such as building/structural monitoring, ground subsidence (due to underground fluid extraction or mining) or inflation (due to build-up of magma below volcanoes), tide gauge stability monitoring, and local tectonic fault motion.

Given the sophistication of scientific GNSS analysis (Chaps. 34, 36, and 37), the computation of positions to subcentimeter accuracy may involve the determination of not only the geodetic coordinates of the GNSS receivers, but also improved estimates of receiver and satellite clock errors, signal biases, GNSS satellite orbits, atmospheric delay biases, and Earth rotation/orientation parameters. The continuous processing of measurements from hundreds of globally distributed CORSs, by a large number of organizations, coordinated by international geodesy initiatives, is a geodetic enterprise that defies easy partitioning into different geodetic surveying applications.

35.2.2 Land Surveying Operations

The goals of GNSS land, engineering, and hydrographic surveying operations are to coordinate many points on the ground, in the air, or on the sea as quickly as possible to the accuracy required by the client and with the coordinate information expressed in relation to a project, map, or geodetic datum.

It is possible to distinguish between three categories of *point coordination*. One is the task of determining coordinates of points or features that exist. Examples include control surveys, detail or topographic surveys, surveys of buildings and land boundaries, built struc-

Table 35.9 Comments on GNSS geodetic surveying methodologies

Field campaign surveys (baseline mode)	<ul style="list-style-type: none"> ● Application: Densification of geodetic control across a local area ● Minimum of a pair of stationary GNSS receivers, operated in single-baseline mode, baseline lengths typically several tens of kilometers ● Conventional static positioning (Sect. 35.1.1), with observation sessions ranging from an hour to many hours ● Considerable redundancy through multiple occupations of monumented ground control points ● Single-baseline measurement processing using commercial software in post-processing mode ● Network solution through secondary adjustment of baselines, with known control point constraints applied to ensure consistency and connection to surrounding geodetic control, see Fig. 35.1.
Field campaign surveys (multistation mode)	<ul style="list-style-type: none"> ● Applications: <ul style="list-style-type: none"> – Establishment of a primary geodetic datum network across a large (national or continental) area – Densification of datum across hundreds of kilometers – Rapid datum maintenance geodetic surveys after major earthquake – Multi-campaign GNSS surveys to detect small land or ice movement over periods of years ● Multiple receivers deployed across a network of monumented ground control points, in a multi-session mode, to ensure that all control points are occupied by GNSS receiver at least once (and ideally twice, or more) ● Static observation session lengths typically from several hours to 24 h (or even longer) ● Data post-processing options: <ul style="list-style-type: none"> – Simultaneous processing of all observed data files in scientific software, with datum constraints applied directly (i. e., ITRF coordinates of some control points) and perhaps also indirectly (use of precise IGS satellite orbit products) – the most rigorous approach – Use web-based processing services such as AUSPOS [35.8], OPUS [35.6], CRCS-PPP [35.7], etc., that link field surveys with surrounding IGS and/or national CORS – however not as rigorous as simultaneous processing of all campaign data in scenario above
Continuously Operating Reference Stations (CORS)	<ul style="list-style-type: none"> ● Applications: <ul style="list-style-type: none"> – <i>Active</i> geodetic control points realizing national datum; and possibly also supporting commercial RTK/N-RTK services – Part of (national or international) network of observing stations, whose data is used to generate geodetic products such as coordinate time series, satellite orbits or clocks, ionospheric and tropospheric parameters, etc. – Instrumentation primarily intended for monitoring tectonic motion, localized ground or structural deformation, etc. ● Observation data typically streamed to central data or analysis center, where data processing may be carried out in real-time, near-RT, or post-processed mode depending upon application ● CORS density may vary from several tens to several hundred (or even thousand) kilometers ● Data analysis may be via: <ul style="list-style-type: none"> – Scientific software similar to that used for post-processing of field campaign data – Specialized software for RT processing to geodetic modeling standard – Commercial baseline or multi-station software to support RTK/N-RTK operations

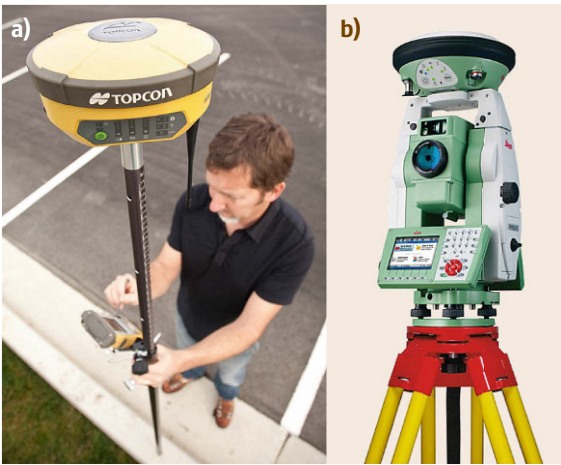
tures, etc. The second is determining the position of a moving object or platform, that is, its trajectory, as in the case of a land vehicle, an aircraft, or a ship. The third is to determine the location of a point that has a prespecified 3-D coordinate – as in set-out surveys on engineering construction sites, or way-points that must be navigated to. The latter two types of positioning are discussed in Sect. 35.3.

GNSS technology intended for use by surveyors and engineers needs to be largely automatic, reliable, and easy-to-operate. For high-productivity operations a commercial off-the-shelf package is preferred – receiver hardware, processing and control software, and ancillary instrumentation. The receiver signal tracking

and processing electronics are essentially identical to geodetic-grade GNSS receivers, and hence are capable of the same measurement quality – although the choke-ring antenna is usually replaced by a light-weight survey antenna (Chap. 17). Furthermore there has been considerable product refinement in survey-grade GNSS receivers, which nowadays are compact, rugged, and come in a variety of form-factors. The most common instrument form-factor is a single unit (without cumbersome antenna or power cabling) containing receiver electronics, antenna, battery, wireless communications, and data memory, able to be placed on a survey pole or other survey instrument (Fig. 35.8), or on a moving platform (Fig. 35.3).

Fig. 35.8a,b Some examples of GNSS receiver form-factors for land surveying applications: (a) pole-mounted GNSS receiver as used by surveyors and engineers to determine coordinates of static points-of-interest (courtesy of Position Partners); (b) GNSS receiver mounted on top of a Total Station supporting integrated survey operation (courtesy of Leica Geosystems) ►

The typical field deployment requires the surveyor’s GNSS receiver to move from one point whose coordinates are to be determined to another, and to continue this procedure until all points have been *visited*. The reference receiver remains set up on the point of known coordinate, and hence the 3-D baseline vectors radiate from that reference station to the points being surveyed, as in Fig. 35.2. This configuration is familiar to land surveyors, as *radiation* is the most common means of determining the coordinates of points – for example, by means of azimuth, distance and vertical angle measure-



ment made from a Total Station instrument (Fig. 35.8) set up on a portable tripod over a fixed ground mark. In

Table 35.10 GNSS land surveying applications and operational issues

Control and deformation surveys	<p>In a continuum of static positioning applications ranging from those that can be identified as geodetic surveying (Table 35.9), to local or project control with the following distinguishing characteristics:</p> <ul style="list-style-type: none">● Generally employ static or rapid-static techniques (Table 35.4), using commercial data-processing software, although RTK/N-RTK mode sometimes used (though with more redundancy and greater care than typical kinematic surveys)● Non-permanent ground marks (e.g., drillholes, nails in kerbs), Datum typically construction project-based, although linked to the national datum at epoch of observation if using RTK/N-RTK● Purpose is closely tied to nearby engineering or surveying activity● Project extent is typically a few to tens of kilometers across● Deformation surveys are associated with construction activities, or focused on built structures, and require either continuous surveys (or re-surveys at regular intervals) of critical points
Topographical surveys and mapping	<p>The rapid determination of the coordinates of many natural surface points or constructed features, across a comparatively small area, with the following characteristics:</p> <ul style="list-style-type: none">● <i>Direct</i> point coordination by GNSS, using rapid-static, stop-&-go, or kinematic surveying techniques (Sect. 35.1)● <i>Indirect</i> mapping, where GNSS is used to determine the precise coordinates of a mapping sensor such as a digital camera or laser scanner● Areal extent is a few hundred square meters to tens (and perhaps hundreds) of square kilometers● Results not required in real-time, though RTK/N-RTK surveys may have lower operational costs● Results may be presented in variety of forms suitable for import into computer aided design (CAD) or geographic information system (GIS) software
Cadastral surveys	<p>Cadastral surveys address legal questions such as: where are the boundaries of land parcel, what rights and responsibilities are attached to a land parcel, and the creation of new land titles following subdivision or redevelopment, and hence have the following characteristics:</p> <ul style="list-style-type: none">● Due to the considerable variety of national and state land titling and cadastral boundary systems, guidelines on what surveyors must measure, to what accuracy, and what information must be registered, will also vary with national or state jurisdiction● There is considerable scope for use of GNSS surveying techniques for cadastral surveys of rural properties; use in urban areas is more problematic● There are very few coordinate-based cadastres, hence GNSS-derived coordinates must be transformed into distances and bearings to be useful for cadastral mapping applications● Survey project extent is typically a few hundreds of meters to perhaps a few kilometers across

the GNSS configuration, the surveyor may not even be responsible for the operation of the reference receiver, and is merely using the RT corrections or the recorded data files (in the case of post-survey computations). Although the algorithms underlying Network-RTK take advantage of data from a network of CORSs, as far as the user is concerned the *packaging* of N-RTK messages is such that it mimics the RT processing of a baseline radiating from a nearby CORS to the user receiver (Table 35.7).

GNSS land and engineering surveying procedures are typically prescribed in national or state standards and recommendations, or contract guidelines, especially for cadastral surveys or datum control surveys (Table 35.10). These standards or guidelines may *suggest*, *recommend*, or *define* the hardware requirements, field observation procedures, ground mark design, quality assurance processes, and minimum and maximum thresholds for geometric constraints such as baseline lengths, network quality checks, number of tracked satellites, and so forth. Nowadays, because GNSS is an all-weather system available 24 h a day that does not require intervisibility between survey receiver and reference receiver(s), there is no longer the need to plan for the best time of day to conduct surveys so as to ensure adequate satellite geometry, or to carry out detailed reconnaissance of the survey area. It is beyond the scope of this chapter to delve into national standards or recommendations for GNSS land, engineering, and hydrographic surveying applications; however the reader is referred to documents such as [35.47–53].

35.2.3 Land Surveying and Mapping Applications

The range of land surveying and mapping applications is very broad (Table 35.10). However, GNSS is but one technology in the land surveyor's toolkit, best suited to clear sky view conditions that ensure that measurements can be made to as many GNSS satellites (with favorable geometry) as possible, and where the *raison d'être* is the determination of position (i. e., coordinates). While the former is a constraint on the operating environment, the latter acknowledges there is a broader set of survey services than just point coordination, which include azimuth or alignment determination, horizontal or vertical offset measurement, and precise physical height (difference) measurement.

The complexity of land and engineering survey tasks requires professional judgement: (a) to select the appropriate technology and operational techniques, (b) to conduct or oversee the careful execution of the field survey, (c) to process measurements taking into account all errors and constraints, and (d) to generate the out-

puts required by the client. The reader is referred to land survey texts for details concerning surveying principles, technologies, and applications [35.54, 55].

Control surveys are similar to geodetic surveys (Sect. 35.2.1); however they are carried out at local or construction project scales [35.54, 55]. The objective is to determine the coordinates of ground control points referred to a project, mapping, or geodetic datum during a field survey campaign. These control points may be of a temporary nature, intended only to be used over a project lifetime, or established as permanent marks (Fig. 35.9). The control points would typically be used for subsequent project surveys for guiding construction, mapping terrain and structures, lower order surveys, or for monitoring ground or structural deformation.

Deformation surveys are a form of geodetic surveying in which the displacement of a GNSS receiver, relative to some *rest* position or position at some measurement epoch, is monitored over time (Sect. 35.2.1). The receiver may be mounted on a deforming engineered structure [35.56], or it may be set up on ground marks in areas of ground surface movement. The distinctions between geodetic deformation surveys and land deformation surveys are largely of a semantic nature; however there are a number of deformation survey scenarios that could be used to distinguish between geodetic, land, and engineering deformation survey applications. It is common to partition those deformation surveys that are sensitive to geophysical or natural processes, such as tectonic motion, volcanic activity, land uplift, or subsidence, from those that measure displacement of engineered structures or monitor deformation with anthropogenic sources such as underground fluid



Fig. 35.9a,b Establishing coordinates of control marks using GNSS: (a) setting up a GNSS receiver/antenna set up over rural control mark using a bipod; (b) GNSS receiver/antenna mounted on tripod to collect measurements for establishing geodetic control at a mine site (courtesy of Position Partners)

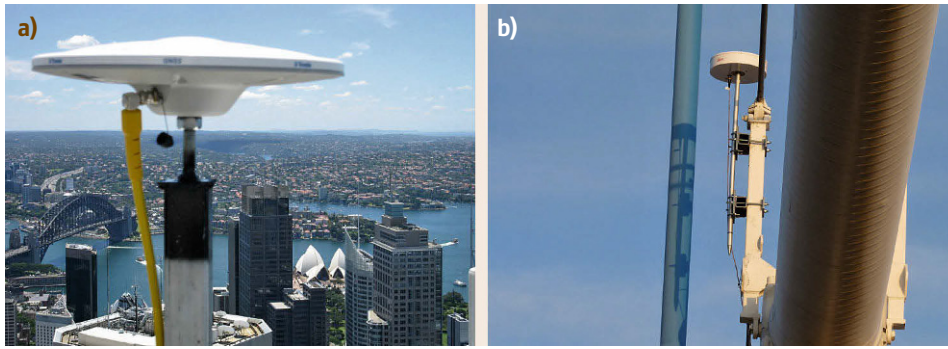


Fig. 35.10a,b GNSS installed on structures for displacement measurements: **(a)** on a tall building in Sydney, Australia (courtesy of Ultimate Positioning); **(b)** attached to support cables of the Severn Suspension Bridge, connecting Bristol to South Wales, UK (courtesy of Gethin Wyn Roberts & Chris J. Brown)

extraction and mining. The former may require more permanent monitoring systems, whereas the latter imply periodic measurement campaigns or monitoring that takes place only over a limited period of time.

Figure 35.10 shows two examples of GNSS installations for deformation monitoring – one on a tall building and the other on a cable suspension bridge. The mode of GNSS positioning may be continuous or episodic, and typically requires some form of time series analysis of computed coordinates in order to detect trends in changes in the coordinates or to determine spectral signatures of vibrating receivers. In addition, GNSS may only be one of a number of technologies that are used in such applications. Other instrumentation include inclinometers and accelerometers.

Topographical surveys are sometimes referred to as *detail surveys*, and are examples of small-area mapping [35.54, 55]. They are similar to surveys carried out using terrestrial survey technology such as Total Stations, except that line-of-sight visibility between reference point and survey point is not necessary. During such surveys the coordinates of ground features (natural and engineered) are determined, including the assumed locations of buried utilities, as well as sufficient sampled surface points to allow the terrain undulations to be modeled as gridded height values, triangulated irregular network points, or contour lines. The output of such surveys is a set of coordinates and feature attributes that permit the data to be exported to CAD or GIS software packages.

Mapping surveys are concerned with the determination of the coordinates of many points across an area for the purpose of describing the terrain, struc-

ture, or built environment in a *spatial* sense [35.54, 57]. Typically what results is a database of coordinates (the *where* information), attributes (the *what* information), and topology (the *how connected* information) of a sufficient number or density of natural or constructed features to ensure a representation of reality at the largest scale of interest. GNSS may be used to directly coordinate the feature to be mapped, or to determine the coordinates of the mapping, imaging, or laser scanning sensor over time, from which coordinates of *pixels* or *point-clouds* are derived in a secondary process.

Cadastral surveys are a special form of survey for the determination or marking-out of land property boundaries [35.58]. In some countries boundaries are defined by coordinates, and hence the survey task is to calculate where the *real* boundaries are with respect to physical structures such as fences, roads, or buildings. However in many countries land boundaries are defined by distance and azimuth of boundary lines as described in registered certificates of titles (in countries that use the Torrens System of title) or in deed documents (for countries that do not) [35.58]. They may also be depicted graphically in cadastral maps. In such cases the GNSS coordinates are used to derive distance and azimuth quantities, and are considered to be one form of evidence that can be used to reconstruct the original land parcel boundaries. GNSS land surveying techniques are particularly useful for rural cadastral surveys, or where new land property boundaries are established as a result of land redevelopment or infrastructure construction projects. [35.48] is an example of GNSS guidelines for cadastral surveying.

35.3 Engineering Surveying

Much of what is stated in Sect. 35.2 with regard to how GNSS is used for land surveying and mapping is also relevant to engineering surveying (see below) and hydrographic surveying (Sect. 35.4). The accuracy requirements are essentially the same, as is the receiver hardware. In addition, there is a reliance on service providers for a variety of augmentation services, and in some cases auxiliary data, to support centimeter-level positioning accuracy. There is an unrelenting drive by GNSS user equipment manufacturers to challenge current operational constraints in order to promote even greater uptake of GNSS technology for engineering applications. Hence some of the most significant innovations are occurring in the GNSS technology that addresses these applications.

Surveys for the construction of roads, bridges, buildings, tunnels, mines, and other structures are based on the same geometric principles and use similar field procedures as land surveying applications [35.53, 54]), and require: (a) the determination of the coordinates of existing ground marks or features, or (b) the identification of marks or points at predefined coordinates to guide construction or machinery. Surveyors are engaged on such projects at all phases of construction, including the original determination of land, building, or marine boundaries. This section focuses on terrestrial engineering applications. Section 35.4 discusses offshore engineering and charting applications.

35.3.1 Engineering Surveying Real-Time Operations

One defining characteristic of almost all GNSS engineering surveying applications is their demand for accurate positioning in real-time. In fact without such capability the application may at best not be cost-effective, or at worse not be feasible at all. RT precise positioning applications include: (a) precise navigation between predefined way-points, such as for vehicle guidance and control applications; (b) construction set-out of formwork, surfaces, and structures; (c) open-cut mining operations; (d) precision agriculture, especially so-called *control track farming*; and (e) rapid mobile mapping. In some cases GNSS is combined with other positioning/guidance technologies – such as laser or vision-based systems, or inertial measurement sensors – to ensure continuous positioning during short GNSS outages, or to provide additional platform orientation information.

RT-GNSS implies no delay between measurements made by the GNSS receiver and the coordinate infor-

mation being generated from measurement processing. Of course there cannot be zero delay; however, it will be assumed that either a delay of one or more seconds is not critical, or computational techniques can be applied to predict position at predefined intervals. RT-GNSS positioning generally implies an *always-on* capability hence the operation of the reference receiver(s) and associated services, such as communications, computing facilities, power, etc., must be continuous, because in addition to high accuracy there is also an increased demand for high *integrity* – machine or vehicle guidance require reliable coordinate solutions.

The flexibility of RT-GNSS positioning is greatest when industry standards for data message transmission, such as those defined by RTCM (Annex A.1.3), are adopted, enabling GNSS receivers from different manufacturers to operate together using the same over-the-air transmissions. The value of industry standards is most obvious in RTK or N-RTK operations (Sect. 35.1.4).

The central role played by RT-GNSS service providers must be acknowledged. Some of the reasons why many users these days take advantage of RT-GNSS services are: (a) the need for continuous and reliable RT operations; (b) the widespread adoption of RTCM data transmission formats; and (c) the high cost/complexity of operating reference receivers. RT-GNSS service providers include private companies, academia, research institutions, and government agencies.

CORSs installed by RT-GNSS service providers typically consist of geodetic-grade receivers, with choke-ring antennas, capable of making multi-frequency, multi-GNSS measurements. RT-GNSS places considerable demands on communication links; between individual CORS receivers and (typically) a central server to manage the transmission of CORS measurements, and for transmission of correction messages to RT-GNSS users. Furthermore, the reference receivers should have the tracking capability to at least match that of the most sophisticated user receiver in order that, for example, RTCM data messages for all visible GNSS satellites and signals that could be used can be broadcast to users. Besides, CORS positioning infrastructure may be used to support GNSS geodesy applications (Sect. 35.2.1) – however the monument on which the GNSS antenna is fixed must be stable.

In contrast to multi-purpose or commercial CORS networks referred to above, there are many RTK systems installed by individual user/operators, especially in the precision agriculture and open-cut mining user segments. These users own several survey-grade GNSS receivers, operate one as a base station, install the other(s) on one (or more) agricultural or mine vehi-

Table 35.11 GNSS engineering surveying applications and operational issues

Construction surveys	<p>Construction surveys support engineering and infrastructure projects, and have the following characteristics:</p> <ul style="list-style-type: none">● GNSS is one technology used by engineers and surveyors on building/construction sites; other technologies must also be used and hence a seamless transfer of coordinates between different construction survey instrumentation is necessary● The immediacy of the tasks on building/construction sites demands the use of real-time GNSS positioning techniques such as RTK/N-RTK; and with increased automation of construction processes there is the need to ensure integration of all types of GNSS positioning on construction sites● Variety of positioning challenges, ranging from coordinating fixed points (similar to topographical surveys), determining coordinates of moving GNSS receiver trajectory, to navigating GNSS receiver to a predefined spatial coordinate● Construction project datum is used, typically requiring the transformation of RTK/N-RTK generated coordinates into the project datum● Survey extent is typically a few hundreds of meters to perhaps a few kilometers across
Construction and mining machinery automation	<p>There is a trend to increased automation of construction and mining machinery automation, from human-in-the-loop implementations to full autonomous operation, implying:</p> <ul style="list-style-type: none">● High-accuracy and high-integrity real-time GNSS positioning availability● Centimeter-level accuracy, though with backup technology options when GNSS is unavailable● Positioning typically is intended to navigate the vehicle to coordinated points, hence requiring adjustment of the vehicle's state from <i>current</i> coordinates to <i>target</i> coordinates● Tight integration with guidance or control systems, and hence often factory-installed by the machine manufacturers themselves● Coverage areas typically up to a few kilometers across
Agriculture	<p>Similar to construction machinery automation applications, with the following unique characteristics:</p> <ul style="list-style-type: none">● Coverage areas may be many kilometers across● The conditions for RT-GNSS are typically more favorable with respect to sky visibility conditions● The accuracy requirements may be more relaxed, ranging from meter-level for standard precision agriculture, to subdecimeter-level accuracy in the case of <i>control track farming</i>● Horizontal positioning
Mapping	<p>Mobile mapping applications are characterized by:</p> <ul style="list-style-type: none">● Variety of platforms – terrestrial, airborne, marine● Variety of mapping sensor technologies, ground sampling (or resolution), field-of-view, cost, operational constraints (e.g., height, range, speed, etc.)● Accuracy requirements may be relaxed considerably, depending upon the mapping methodology that is used● Real-time positioning is in general not essential● Survey extent may vary from a few kilometers to many tens of kilometers across

cle, and implement a *closed* RTK service via a UHF radio link between the receivers. Such a configuration is not optimal, for no other reason than the wasteful duplication of base stations across a coverage area. It is expected that, over time, such GNSS users will decommission their own base stations and instead subscribe to RT-GNSS services.

35.3.2 Engineering Surveying Applications

These type of surveys may be considered a subcategory of GNSS *land surveying applications* (Sect. 35.2.3), and are those that are: (a) undertaken on land, (b) associated with construction or mining activities, (c) limited to a project area, (d) constrained to a project time scale,

(e) involve machinery, and (f) extensively use RT-GNSS techniques. Examples of engineering surveying applications are listed in Table 35.11, and discussed below.

Construction surveys address the different positioning requirements of civil engineers and building professionals during the construction phase for any engineered structure [35.54]. High-accuracy GNSS technology is used in place of traditional terrestrial surveying instrumentation for setting out of trenches or formwork for concrete pours, checking verticality (or horizontality) of construction, or measuring the dimensions of structural component, such as walls, beams, pipes, cables, and so on (Fig. 35.11). The utility of being able to do this in real-time is crucial in order that immediate action can be taken, whether in the form of routine execution of



Fig. 35.11 GNSS as typically used by surveyors on construction sites: here is shown a pole-mounted GNSS receiver which is being used to either coordinate a point-of-interest or to mark a point whose coordinate is provided in order to set out formwork for concrete pouring, laying of cables, pipes, or services, etc.; typically operating in real-time mode (courtesy of Leica Geosystems)

engineering tasks or to allow for on-site modification or adjustment of construction plans. These surveys are, in many respects, the most demanding applications of high-accuracy GNSS technology because of the variable conditions on construction sites. For example, there may be significant shading of the sky, considerable vehicular and human traffic, dangerous/noisy/dirty conditions, variable wireless coverage, and a number of different coordinate datums, to name but a few. The engineering surveyor must be capable of executing their tasks in an often stressful and unpredictable environment. Furthermore, GNSS is but one tool at their disposal. However, there is a trend to increased *automation* of excavating, drilling, concreting, paving, laying of preformed slabs, erection of walls or formwork, removal of waste material, etc., which implies instantaneous guidance and/or control of heavy machinery using high-accuracy, high-integrity GNSS technology, possibly supplemented with laser, vision, and inertial systems to improve availability and reliability.

Construction machinery automation of graders, bulldozers, tractors, trucks, and specialized vehicles or machinery brings with it improvements in productivity [35.59]. This productivity can be measured in many ways, including faster and more accurate construction, longer work days, with fewer errors, smaller construction workforce, less injuries to workers, and reduced fuel use. Early examples of machine automation for construction environments are closely related to the technology supporting precision agriculture, especially *control traffic farming* [35.60, 61] where RT-GNSS is used to guide farm machinery with an accuracy that ensures the vehicle's wheel ruts are always in the



Fig. 35.12a,b GNSS receivers installed on construction machinery to guide excavations (a) and grading (b). Note in (a) that two antennas/receivers are installed to allow for GNSS to determine not just position, but also orientation in 3-D so that the bulldozer blade may be manipulated to excavate an inclined design surface ((a) courtesy of Leica Geosystems, (b) courtesy of Ultimate Positioning)

same *track*. This requires subdecimeter, repeatable positioning accuracy, in real-time on a continuous basis. Construction equipment can be similarly *guided* along *tracks*, ensuring road centerlines are set out according to design coordinates, or the concreting of airport runways and taxiways is carried out very precisely in a vertical sense. This is in contrast to kinematic GNSS positioning as illustrated in Fig. 35.3, where the GNSS instrument is used to map the terrain as it actually is. Figure 35.12 shows GNSS receivers mounted on construction machinery.

Over the next decade the degree of autonomy of construction vehicles and machinery will increase significantly, and construction, mining, and agriculture will likely be the largest markets for high-accuracy GNSS positioning systems. Machine automation can be implemented in a variety of scenarios, from simply aiding the operator via in-cabin computer displays that show actual vehicle tracks and design lines or surfaces (Fig. 35.13) through radio-controlled machinery by operators who may not even be located on the site to fully autonomous robots that operate with no human intervention at all. These applications require centimeter-level positioning accuracy provided by RT-GNSS. However, the level of integrity may range from relatively low – with the operator merely informed when positioning is unavailable, who then controls the machinery manually – to very high integrity in the case of full machine automation. Yet even in this mode the addition of vision or scanning sensors can provide enough *situation-awareness* for an autonomous vehicle to respond to loss of GNSS positioning capability.



Fig. 35.13 Inside cabin of GNSS-assisted machinery – one or more GNSS receivers/antennas (multiple antennas provide vehicle orientation information) are installed on construction machinery and real-time solutions for position (and perhaps orientation angles) of the reference point on the vehicle are displayed to the machinery operator on a controller device together with the planned trajectory of the machinery so that the excavation may be carried out according to design (courtesy of Ultimate Positioning)

Mining survey applications are subcategories of several GNSS land and engineering surveying applications [35.54, 55]. It must first be acknowledged that GNSS can only be used for open-cut mine operations, which are similar to construction project sites. On such sites the full range of surveying and positioning tasks are required: mapping, set-out, construction, volume surveys, machine guidance/control, and vehicle fleet management/tracking. As with construction site GNSS surveys, the area of operations is rather constrained – perhaps just a few kilometers across – and the dirty, dangerous, and typically extreme environmental conditions place heavy demands on technology. The challenge for RT-GNSS users in deep open-cut mines is that with increasing depth, the proportion of open sky that a GNSS receiver *sees* decreases rapidly. This is especially the case when surveyors or GNSS-guided machinery are working near steeply sloping mine walls. It was the need to increase the number of visible satellites under such conditions, beyond the available GPS constellation, which has driven the adoption of multi-GNSS receivers for such critical applications – initially GPS+GLONASS, but nowadays capable of tracking signals, and processing measurements, from other GNSS constellations.

35.3.3 Project Execution and Related Issues

The applications listed above imply operations over relatively small areas. GNSS must compete with terrestrial

surveying technologies, and hence must be a cost-effective and easy-to-use technology. It should be used only in project environments that are optimal for rapid and reliable AR, and for which there is very good sky visibility. Unlike GNSS geodetic or land surveying projects, reconnaissance *prior* to the use of GNSS for engineering surveys is not carried out.

In addition, given the construction project scale of most engineering surveying applications, the issue of the coordinate datum is different to that for geodetic surveying, and perhaps even to land surveying. The datum is typically of local relevance, with coordinates often expressed in a horizontal map projection for ease of graphical display and spatial analysis. The vertical component is measured in terms of physical heights, not ellipsoidal heights (or height differences). Hence *horizontal surveys* are typically carried out, with *vertical surveys* often conducted using one of a number of lev-



Fig. 35.14 Mobile mapping system (MMS) installed on a road vehicle. The system comprises multiple imaging sensors (cameras pointing forward, sideways, and backward), a laser scanner (on the top of the vehicle), a GNSS antenna (at the top of the van), and an inertial navigation system for platform orientation (box on rack next to laser scanner). Note also that this particular MMS is carrying additional sensors (mounted low to the ground) for radar imaging of the road surface and detection of cracks in pavement (courtesy of Charles Toth)

eling techniques, including the GNSS ellipsoidal height + geoid height-leveling method [35.62].

However, RTK/N-RTK operations imply that coordinates are determined in the datum defined by the coordinates of the CORSS – which typically are expressed in a national reference frame. In some instances, for example, at a mine, dam, and other large construction site the reference receivers are operated by the project surveyors, and the RTK/N-RTK settings may be adjusted to output GNSS coordinates in the local project datum or coordinate system. The situation regarding RT-PPP is more complex as the point positioning technique derives its datum from the precise GNSS satellite orbits, and these are invariably in a globally relevant, stable reference frame such as the ITRF [35.18]. In summary, for RT applications GNSS-derived coordinates may need to be transformed into the local project datum by the GNSS field instrumentation before they can be used by engineering surveyors, or by the machines that are guided by RT-GNSS systems.

With respect to GNSS-enabled *mapping* (Sect. 35.2.3), although the mapped points may be station-

ary, these days geospatial data acquisition is carried out from a moving platform (e.g., equipped with a camera or laser scanner) such as a vehicle (Fig. 35.14), aircraft (or unmanned aerial vehicle), or ship. Maximum flexibility is afforded by *post-survey* processing of recorded GNSS measurements. In addition, decimeter-level or lower accuracy is typically adequate allowing for relaxed instrument or field operational requirements. Furthermore the 3-D orientation or attitude of the mapping sensor is typically determined using inertial technology (Chap. 28). The operational guidelines, quality control procedures, and accuracy requirements for different mobile mapping platforms will vary considerably. It is beyond the scope of this chapter to discuss in detail the range of mobile mapping applications, the imaging and scanner technologies that are available, the mapping analysis methodologies that can be used, and the operational guidelines to be followed. Readers are referred to [35.63], and similar articles in geospatial magazines and international conference proceedings, for the latest developments in this rapidly evolving technological field.

35.4 Hydrographic Surveying

Much of what is stated in Sect. 35.3 with regard to how GNSS is used for engineering surveying and mapping is also relevant to hydrographic surveying in support of offshore engineering and sea floor charting. Offshore engineering associated with pipelines, undersea cables, breakwaters, harbor works, and free-standing structures has similar requirements for pre-construction surveys; for subsequent support or control of operations during the construction phase; and, finally, postconstruction *as-built* surveys. Furthermore, charting surveys require positioning of the moving platform, similar to terrestrial or aerial mapping, although the undersea imaging technologies are very different.

35.4.1 Hydrographic Surveying Applications

Although land, engineering, and offshore surveying share many geometric principles [35.53], the offshore operational environment is in many respects more challenging [35.64]. The environment is more corrosive, the marine platform (such as a ship, drill-rig, dredging vessel, small boat, or autonomous underwater vehicle) is in continuous motion, and the distances from marine receivers to shore-based reference stations may be longer than is the case for most land-based applications. On the other hand sky visibility is typically very good.

Prior to the introduction of GNSS, the techniques for offshore positioning were less accurate, more complex, and more expensive than those used on land. Invariably as the distance from shore increased, the positioning accuracy reduced, and the *electronic* positioning technology that could be used changed. The positioning technology was classified as *short-range*, *medium-range*, or *long-range*, referring to the distance over which transmitted terrestrial ranging signals could be detected [35.65, 66]. The introduction of the Transit Navy Navigation Satellite System (also often referred to as *Transit Doppler*) to the civilian community in 1964 [35.67, 68] made it possible to undertake hydrographic survey operations anywhere in the world, without relying on shore-based signal transmitters. The Transit Doppler system was retired in 1996, but GPS further revolutionized hydrographic surveying and maritime navigation. Nowadays, GNSS is used for all (surface) marine positioning requirements [35.66].

Hydrographic surveying operations can be partitioned into two general classes [35.64, 65]: (a) charting and (b) offshore engineering activities (Table 35.12). As with land-based mapping and surveying applications, some require RT positioning while others may be addressed using post-processed techniques.

Charting is an operation in which a mapping sensor aboard a ship, or towed *fish*, moves in a pattern that ensures an entire area of the seabed is imaged, or *illu-*

Table 35.12 GNSS hydrographic surveying and marine applications

Harbor and river operations	<ul style="list-style-type: none">● Typical applications: small-scale surveys of river or harbor bed, positioning of buoys, cables or pipelines, vessel-docking maneuvers, etc.● Scenarios vary: positioning vessel or structure; measuring vessel’s trajectory; navigating to predefined locations● If accuracy demands it, RTK/N-RTK techniques are used
Dredging	<ul style="list-style-type: none">● Similar to land-based engineering surveys, requiring precise real-time spatial positioning of vessel-mounted excavating equipment● Attitude of vessel may be determined using a GNSS multi-antenna system, although using an inertial system is a common option● Typically conducted close to shore, permitting the use of standard RTK techniques
Offshore engineering	<ul style="list-style-type: none">● Typical applications: construction of breakwaters, piers, shore defences, wind or tidal energy platforms, gas and oil-drilling platforms, pipelines, cable laying● Operations will vary from being very close to land, to mid-ocean● Accuracy requirements will vary considerably, hence there is a wide choice of GNSS positioning techniques● Real-time positioning is typically required
Charting	<ul style="list-style-type: none">● Can take place well offshore, for which differential kinematic positioning techniques may be impracticable● Horizontal positioning accuracy is defined by international standards, and rarely requires the use of carrier-phase-based techniques● Chart Datum is typically lowest astronomical tide, hence vertical (ellipsoidal) positioning of sonar sensor not required, although the vessel’s heave motion is measured so as to correct raw depth measurements● Real-time positioning is rarely a requirement for charting

minated, by transmitted sound waves, and the reflected signals recorded – the acoustic sensor may be a side-scan sonar or an echo sounder [35.64, 69]. Much like an airborne or vehicle-mounted camera or laser scanner, the return signals are processed to generate a 3-D map of the (reflecting) surface (Fig. 35.15). As with other types of mapping, both the *position* and *orientation*



Fig. 35.15 Multibeam sonar used to derive digital elevation model of seabed requires the position and orientation of sonar sensor attached to survey vessel so as to convert range measurements into coordinates of reflecting surface, which may be transformed into electronic nautical charts for navigation or to support offshore engineering (courtesy of Spain Hydrographic Service)

of the sensor must be measured so that direct georeferencing techniques can be used. In the case of active mapping systems such as sonar or laser scanners, the position and orientation of both the signal transmitter and the signal receiver are required, while this requirement needs to be fulfilled only for the imaging sensor. Unlike land GNSS applications, for which there are no internationally recognized standards and recommendations on how to execute GNSS surveys, charting operations follow guidelines such as those from the International Hydrographic Organization (IHO; [35.70, 71]).

Surveys in support of *offshore engineering* are similar to construction surveys (Sect. 35.3). Offshore con-



Fig. 35.16 Positioning of offshore cable laying ships and drill platforms is nowadays undertaken using GNSS technology (courtesy of Alf van Beem, after [35.72])

struction applications use identical GNSS surveying instrumentation and techniques to land-engineering surveys. GNSS is used to guide the placement of undersea pipelines or cables (Fig. 35.16), or the erection of offshore structures such as drill platforms, wind or tidal energy generating turbines, or breakwaters, and other river, harbor, or open ocean works. *Dredging*, for example, requires similar technology as does operator guidance of construction machinery (Fig. 35.13), as the objective is to excavate a channel, river, or portion of the seabed to some desired depth.

35.4.2 Operational Issues

There are several unique characteristics of hydrographic surveying worthy of mention. While high-accuracy marine positioning is still based on differential positioning principles, the challenge of operating well offshore, at long distances from GNSS reference stations, means that there is greater interest in using alternative high-accuracy positioning techniques for offshore positioning than is the case on land. Hence the offshore positioning market is an early adopter of PPP techniques, with several service providers transmitting satellite orbit and clock information to support RT-PPP [35.73–76].

Although many hydrographic and charting surveys are carried out near to shore, and even within a harbor, the coordinate datum is in general different to the land geodetic datum. Many offshore engineering surveys use a project datum (as do many onshore engineering projects). The IHO has mandated that the horizontal datum for all charting must be that of WGS84 [35.69, 77] – for all intents and purposes an ITRF-aligned datum. The guidelines for hydrographic surveys also tend to be internationally applicable [35.70, 71]. The RT-GNSS service providers who cater for the offshore surveying market are companies that operate on a global basis.

The seabed map is typically used for ship navigation, hence all underwater or exposed obstacles that pose a danger to maritime shipping should be accurately surveyed. There must be under-keel clearance of the traversing vessel, hence the vertical, or depth accuracy is required to be higher than the horizontal accuracy of any map feature. For all but the largest scale charts this implies a horizontal accuracy of no better than 5–10 m (and often much worse), while the depth accuracy requirement in rivers, harbors, and shipping channels may be at the submeter-level (and often higher). The *chart datum* is typically the Lowest Astronomical Tide [35.78].

References

- 35.1 C. Rizos: Making sense of the GNSS techniques. In: *Manual of Geospatial Science and Technology*, 2nd edn., ed. by J. Bossler, J.B. Campbell, R. McMaster, C. Rizos (Taylor Francis, London 2010) pp. 173–190
- 35.2 C. Rizos, D. Grejner-Brzezinska: GPS positioning models for single point and baseline solutions. In: *Manual of Geospatial Science and Technology*, 2nd edn., ed. by J. Bossler, J.B. Campbell, R. McMaster, C. Rizos (Taylor Francis, London 2010) pp. 135–149
- 35.3 A. Leick, L. Rapoport, D. Tatarnikov: *GPS Satellite Surveying*, 4th edn. (Wiley, Hoboken 2015)
- 35.4 B. Hoffmann-Wellenhof, H. Lichtenegger, E. Wasle: *GNSS – Global Navigation Satellite Systems* (Springer, Wien, New York 2008)
- 35.5 RINEX – The Receiver Independent Exchange Format – Version 3.02 3 Apr. 2013 (IGS RINEX WG and RTCM-SC104, 2013)
- 35.6 NGS: National Geodetic Survey's (NGS) OPUS web processing site. <http://www.ngs.noaa.gov/OPUS/>
- 35.7 Natural Resources Canada (NRCAN): Canadian Spatial Reference System Precise Point Positioning (CSRS-PPP) web processing site <http://webapp.geod.nrcan.gc.ca/geod/tools-outils/ppp.php?locale=en>
- 35.8 Geoscience Australia: AUSPOS online GPS processing service. <http://www.ga.gov.au/scientific-topics/positioning-navigation/geodesy/auspos/>
- 35.9 U. Vollath, H. Landau, X. Chen, K. Doucet, C. Pagels: Network RTK versus single base RTK – Understanding the error characteristics, Proc. ION GPS 2002, Portland (ION, Virginia 2002) pp. 2774–2781
- 35.10 RTCM Standard 10403.2 Differential GNSS Services, Version 3 with Amendment 2, 7 Nov. 2013 (RTCM, Arlington 2013)
- 35.11 G. Wübbena, A. Bagge, G. Seeber, V. Boder, P. Hanckemeier: Dependent errors for real-time precise DGPS applications by establishing stations networks, Proc. ION GPS 1996, Kansas City (ION, Virginia 1996) pp. 1845–1852
- 35.12 L. Wanning: Real-time differential GPS error modelling in regional reference station networks. In: *Advances in Positioning and Reference Frames*, International Association of Geodesy Symposia, Vol. 118, ed. by F.K. Brunner (Springer, Berlin 1998) pp. 86–92
- 35.13 L. Dai, S. Han, J. Wang, C. Rizos: A study on GPS/GLONASS multiple reference station techniques for precise real-time carrier phase-based positioning, Proc. ION GPS 2001, Salt Lake City (ION, Virginia 2001) pp. 392–403
- 35.14 G. Fotopoulos, M.E. Cannon: An overview of multi-reference station methods for cm-level positioning, GPS Solutions 4(3), 1–10 (2001)

- 35.15 H. Landau, U. Vollath, X. Chen: Virtual reference station systems, *J. Glob. Position. Syst.* **1**(2), 137–143 (2002)
- 35.16 C. Rizos: Network RTK research and implementation: A geodetic perspective, *J. Glob. Position. Syst.* **1**(2), 144–150 (2002)
- 35.17 S. Hilla: Extending the standard product 3 (SP3) orbit format, *Proc. Int. GPS Serv. Netw. Data Anal. Center Workshop*, Ottawa (IGS, Pasadena 2002)
- 35.18 Z. Altamimi, X. Collilieux, L. Métivier: ITRF2008: An improved solution of the international terrestrial reference frame, *J. Geod.* **85**(8), 457–473 (2011)
- 35.19 J.F. Zumberge, M.B. Hefflin, D.C. Jefferson, M.M. Watkins, F.H. Webb: Precise point positioning for the efficient and robust analysis of GPS data from large networks, *J. Geophys. Res.* **102**(B3), 5005–5017 (1997)
- 35.20 P. Héroux, Y. Gao, J. Kouba, F. Lahaye, Y. Mireault, P. Collins, K. Macleod, P. Tetreault, K. Chen: Products and applications for precise point positioning – Moving towards real-time, *Proc. ION GPS 2004*, Long Beach (ION, Virginia 2004) pp. 1832–1843
- 35.21 R.J.P. van Bree, C. Tiberius: Real-time single-frequency precise point positioning: Accuracy assessment, *GPS Solutions* **16**(2), 259–266 (2012)
- 35.22 C. Tiberius, R. van Bree, P. Buist: Staying in lane – Real-time single-frequency PPP on the road, *Inside GNSS* **6**(6), 48–53 (2011)
- 35.23 H. van der Marel, P. de Bakker: Single versus dual-frequency precise point positioning, *Inside GNSS* **7**(4), 30–35 (2012)
- 35.24 H.J. Euler, C.R. Keenan, B.E. Zebhauser, G. Wübbena: Study of a simplified approach in utilizing information from permanent reference station arrays, *Proc. ION GPS 2001*, Salt Lake City (ION, Virginia 2001) pp. 379–391
- 35.25 B.E. Zebhauser, H.J. Euler, C.R. Keenan, G. Wübbena: A novel approach for the use of information from reference station networks conforming to RTCM V2.3 and future V3.0, *Proc. ION NTM 2002*, San Diego (ION, Virginia 2002) pp. 863–876
- 35.26 F. Takac, O. Zelzer: The relationship between network RTK solutions MAC, VRS, PRS, FKP and i-MAX, *Proc ION GPS 2008*, Savannah (ION, Virginia 2008) pp. 348–355
- 35.27 J. Kouba, P. Héroux: Precise point positioning using IGS orbit and clock products, *GPS Solutions* **5**(2), 12–28 (2001)
- 35.28 S. Bisnath, Y. Gao: Current state of precise point positioning and future prospects and limitations. In: *Observing Our Changing Earth*, International Association of Geodesy Symposia, Vol. 133, ed. by M. Sideris (Springer, Berlin, Heidelberg 2009) pp. 615–623
- 35.29 S. Bisnath, P. Collins: Recent developments in precise point positioning, *Geomatica* **66**(2), 103–111 (2012)
- 35.30 M. Caissy, L. Agrotis, G. Weber, M. Hernandez-Pajares, U. Hugentobler: Coming soon – The international GNSS real-time service, *GPS World* **23**(6), 52 (2012)
- 35.31 International GNSS Service (IGS) Real-Time Service (RTS) web site. <http://igs.org/rtts>
- 35.32 O. Øvstedal: Absolute positioning with single-frequency GPS receivers, *GPS Solutions* **5**(4), 33–44 (2002)
- 35.33 Y. Gao, Y. Zhang, K. Chen: Development of a real-time single-frequency precise point positioning system and test results, *Proc. ION GNSS 2006*, Fort Worth (ION, Virginia 2006) pp. 2297–2303
- 35.34 A.Q. Le, C. Tiberius: Single-frequency precise point positioning with optimal filtering, *GPS Solutions* **11**(1), 61–69 (2007)
- 35.35 G. Wübbena, M. Schmitz, A. Bagg: PPP-RTK: Precise point positioning using state-space representation in RTK networks, *Proc. ION GNSS 2005*, Long Beach (ION, Virginia 2005) pp. 2584–2594
- 35.36 L. Mervart, Z. Lukes, C. Rocken, T. Iwabuchi: Precise point positioning with ambiguity resolution in real-time, *Proc. ION GNSS 2008*, Savannah (ION, Virginia 2008) pp. 397–405
- 35.37 M. Ge, G. Gendt, M. Rothacher, C. Shi, J. Liu: Resolution of GPS carrier-phase ambiguities in precise point positioning (PPP) with daily observations, *J. Geod.* **82**(7), 389–399 (2008)
- 35.38 S. Loyer, F. Perosanz, F. Mercier, H. Capdeville, J.-C. Marty: Zero-difference GPS ambiguity resolution at CNES-CLS IGS Analysis Center, *J. Geod.* **86**(11), 991–1003 (2012)
- 35.39 P.J.G. Teunissen, A. Khodabandeh: Review and principles of PPP-RTK methods, *J. Geod.* **89**(3), 217–240 (2015)
- 35.40 T. Sagiya: A decade of GEONET: 1994–2003 – The continuous GPS observation in Japan and its impact on earthquake studies, *Earth Planets Space* **56**(8), xxix–xlii (2004)
- 35.41 D. Norin, J. Sunna, R. Lundell, G. Hedling, U. Olsson: Test of RTCM version 3.1 network RTK correction messages (MAC) in the field and on board a ship for uninterrupted navigation, *Proc. ION GNSS 2012*, Nashville (ION, Virginia 2012) pp. 1147–1157
- 35.42 C. Bruyninx: The EUREF permanent network: A multi-disciplinary network serving surveyors as well as scientists, *Geoinformatics* **7**(5), 32–35 (2004)
- 35.43 J.M. Dow, R.E. Neilan, C. Rizos: The international GNSS service in a changing landscape of global navigation satellite systems, *J. Geod.* **83**(3/4), 191–198 (2009)
- 35.44 W.R. Dick, B. Richter: The International Earth Rotation and Reference Systems Service (IERS). In: *Organizations and Strategies in Astronomy*, Vol. 5, ed. by A. Heck (Kluwer Academic, Dordrecht 2004) pp. 159–168
- 35.45 H.-P. Plag, M. Pearlman (Eds.): *Global Geodetic Observing System: Meeting the Requirements of a Global Society on a Changing Planet in 2020* (Springer, Berlin, Heidelberg 2009)
- 35.46 T. Herring: *Geodesy: Treatise on Teophysics*, Vol. 3 (Elsevier, New York 2009)
- 35.47 Guideline for Control Surveys by GNSS, Special Publication 1, v.2.1 (Australia's Intergovernmental Committee for Surveying and Mapping, Canberra 2014)

- 35.48 Guidelines for cadastral surveying using GNSS. In: *Survey Practice Handbook – Part 2: Survey Procedures* (Surveyors Registration Board of Victoria, Melbourne 2006) pp. 1–30
- 35.49 J. Wentzel, B. Donahue, R. Berg: *Guidelines for RTK/RTN GNSS Surveying in Canada, v.1.1* (Natural Resources Canada, Ottawa 2013)
- 35.50 W. Henning: *User Guidelines for Single Base Real Time GNSS Positioning, v3.1.1* (National Oceanic and Atmospheric Administration, National Geodetic Survey, Silver Spring 2011)
- 35.51 F.G.C. Committee: *Geometric Geodetic Accuracy Standards and Specifications for Using GPS Relative Positioning Techniques*, 5th edn. (National Geodetic Survey, NOAA, Rockville 1989)
- 35.52 Guidance Notes for GNSS RTK Surveying in Great Britain, 4th edn. (The Survey Association, Newark-on-Trent 2015)
- 35.53 GPS survey Specifications. In: *Surveys Manual* (California's Department of Transport, Office of Land Surveys, Sacramento 2012)
- 35.54 B.F. Kavanagh, S.J.G. Bird: *Surveying: Principles and Applications*, 9th edn. (Prentice Hall, Upper Saddle River 2013)
- 35.55 J. Uren, W.F. Price: *Surveying for Engineers*, 5th edn. (Palgrave Macmillan, London 2010)
- 35.56 C. Ogaja, X. Li, C. Rizos: Advances in structural monitoring with global positioning system technology: 1997–2006, *J. Appl. Geod.* **1**(3), 171–179 (2007)
- 35.57 K. Kraus: *Photogrammetry: Geometry from Images and Laser Scans* (Walter de Gruyter, Berlin 2007)
- 35.58 P. Dale, J. McLaughlin: *Land Administration* (Oxford Univ. Press, Oxford 2000)
- 35.59 C. Rizos: GPS, GNSS and the future. In: *Manual of Geospatial Science and Technology*, 2nd edn., ed. by J. Bossler, J.B. Campbell, R. McMaster, C. Rizos (Taylor Francis, London 2010) pp. 259–281
- 35.60 G.D. Vermeulen, J.N. Tullberg, W.C.T. Chamen: Controlled traffic farming. In: *Soil Engineering*, ed. by A.P. Dedousis, T. Bartzanas (Springer, Berlin 2010) pp. 101–120
- 35.61 B. Whelan, J. Taylor: *Precision Agriculture for Grain Production Systems* (CSIRO Publishing, Collingwood 2013)
- 35.62 C. Rizos: Carrying out a GPS surveying/mapping task. In: *Manual of Geospatial Science and Technology*, 2nd edn., ed. by J. Bossler, J.B. Campbell, R. McMaster, C. Rizos (Taylor Francis, London 2010) pp. 217–234
- 35.63 G. Petrie: Mobile mapping systems – An introduction to the technology, *GEoinformatics January/February*, 32–43 (2010)
- 35.64 A.E. Ingham, V.J. Abbott: *Hydrography for the Surveyor and Engineer*, 3rd edn. (Wiley-Blackwell, Hoboken 1993)
- 35.65 R.P. Loweth: *Manual of Offshore Surveying for Geoscientists and Engineers* (Chapman Hall, London 1997)
- 35.66 A. Peacock: *The Principles of Navigation: The Admiralty Manual of Navigation*, Vol. 1, 10th edn. (The Nautical Institute, London 2008)
- 35.67 T.A. Stansell: The Navy navigation satellite system: Description and status, *Navigation* **15**(3), 229–243 (1968)
- 35.68 R.J. Danchik: An overview of transit development, *John Hopkins APL Tech. Digest* **19**(1), 18–26 (1998)
- 35.69 Manual on Hydrography, 1st edn. (International Hydrographic Bureau, Monaco 2011)
- 35.70 IHO Standards for Hydrographic Surveying, 5th edn., Special Publication No. 44, (International Hydrographic Bureau, Monaco 2008)
- 35.71 Regulations of the IHO for International Charts and Chart Specifications of the IHO, edn. 4.4.0, Special Publication No. 4, (International Hydrographic Bureau, Monaco 2013)
- 35.72 Photo source: [https://commons.wikimedia.org/wiki/Category:Ndurance_\(ship,_2012\)#/media/File:Ndurance_-_IMO_9632466_leaving_Port_of_Rotterdam,_pic1.JPG](https://commons.wikimedia.org/wiki/Category:Ndurance_(ship,_2012)#/media/File:Ndurance_-_IMO_9632466_leaving_Port_of_Rotterdam,_pic1.JPG)
- 35.73 L. Rodrigo, H. Landau, M. Nitschke, M. Glocker, S. Seeger, X. Chen, A. Deking, M. BenTahar, F. Zhang, K. Ferguson, R. Stolz, N. Talbot, G. Lu, T. Allison, M. Brandl, V. Gomez, W. Cao, A. Kipka: RTX Positioning: The next generation of cm-accurate real-time GNSS positioning, *Proc. ION GNSS 2011*, Portland (ION, Virginia 2011) pp. 1460–1475
- 35.74 T. Melgard, E. Vigen, O. Orpen: Advantages of combined GPS and GLONASS PPP – Experiences based on G2, a new service from Fugro, *Proc. 13th IAIN World Congress*, Stockholm (IAIN, London 2009) pp. 1–7
- 35.75 L. Dai, R.R. Hatch: Integrated StarFire GPS with GLONASS for real-time precise navigation and positioning, *Proc. ION GNSS 2011*, Portland (ION, Virginia 2011) pp. 1476–1485
- 35.76 C. Rocken, L. Mervart, J. Johnson, Z. Lukes, T. Springer, T. Iwabuchi, S. Cummins: A new real-time global GPS and GLONASS precise positioning correction service: Apex, *Proc. ION GNSS 2011*, Portland (ION, Virginia 2011) pp. 1825–1838
- 35.77 Department of Defense World Geodetic System 1984 (WGS84): Its Definition and Relationships with Local Geodetic Systems, Publication NIMA TR8350.2, 3rd ed., amendm. 1 (National Imagery and Mapping Agency, Reston 2000)
- 35.78 Tidal Datums and their Applications, NOAA Special Publication NOS CO-OPS 1 (National Oceanic and Atmospheric Administration, Silver Spring 2000)

Geodesy

36. Geodesy

Zuheir Altamimi, Richard Gross

Continuous geodetic observations are fundamental to characterize changes in space and time that affect the Earth system. The advent of global navigation satellite systems (GNSSs), starting with the Global Positioning System (GPS) in the early 1980s, has significantly increased the range of geodetic applications and their precision. Significant improvements have progressively been made in the GNSS software packages developed by research institutes, leading to the determination of high-precision geodetic parameters and their temporal variations. The proliferation of dense GNSS networks (local, national, continental and global), composed of continuously observing stations, allows for a variety of geodetic and Earth science applications. Most areas of science, Earth observation, georeferencing applications, and society at large, today depend on being able to determine positions to millimeter-level precision. Point positions, to be meaningful and fully exploitable, have to be determined and expressed in a well-defined reference frame. All current global and regional reference frames rely on the availability of the international terrestrial reference frame (ITRF), which is the most accurate realization of the international terrestrial reference system (ITRS). One of the major modern achievements in geodesy today is the ability to determine highly precise global and regional terrestrial reference frames based on GNSS observations, fully connected to the ITRF. This chapter describes the use and applications of GNSS in geodesy, focusing on its role in the International

36.1 GNSS and IAG's Global Geodetic Observing System	1039
36.1.1 The International Association of Geodesy	1040
36.1.2 The Global Geodetic Observing System	1041
36.2 Global and Regional Reference Frames	1044
36.2.1 Reference Frame Representations for the Deformable Earth	1044
36.2.2 Global Terrestrial Reference Frames	1047
36.2.3 GNSS-Based Reference Frames and Their Relationship with the ITRF	1050
36.2.4 General Guidelines for GNSS-Based Reference Frame Implementation	1052
36.2.5 GNSS, Reference Frame and Sea Level Monitoring	1053
36.3 Earth Rotation, Polar Motion, and Nutation	1054
36.3.1 Theory of the Earth's Rotation	1055
36.3.2 Length-of-Day	1055
36.3.3 Polar Motion	1056
36.3.4 Nutation	1058
References	1059

Association of Geodesy's (IAG's) global geodetic observing system (GGOS) for monitoring our planet in space and time, GNSS-based reference frame implementation, Earth rotation and sea level monitoring.

36.1 GNSS and IAG's Global Geodetic Observing System

Geodesy is the science of the Earth's rotation, gravity and shape, including their evolution in time [36.1,2]. These properties of the Earth change in time because the Earth is a dynamic system – it has a fluid, mobile atmosphere and oceans, a continually changing global distribution of ice, snow, and water, a fluid core that is

undergoing some type of hydromagnetic motion, a mantle both thermally convecting and rebounding from the glacial loading of the last ice age, and mobile tectonic plates (Chap. 37). In addition, external forces due to the gravitational attraction of the Sun, Moon, and planets also act upon the Earth. These internal dynamical pro-

cesses and external gravitational forces exert torques on the solid Earth, or displace its mass, thereby causing the Earth's rotation, gravity, and shape to change. Geodetic observing systems, including the space-geodetic techniques of very long baseline interferometry (VLBI), satellite laser ranging (SLR), global navigation satellite systems (GNSSs) like the US Global Positioning System (GPS), and the French Doppler orbitography and radio-positioning by integrated satellite (DORIS) system, provide the measurements of the Earth's rotation, gravity, and shape that are used to study the response of the Earth to these dynamical forces.

Observations of the Earth's variable rotation, gravity and shape also provide the basis for the realization of the reference systems that are required in order to assign coordinates to points and objects and thereby determine how those points and objects move in space and time (Fig. 36.1). The terrestrial reference frame (TRF) determined by geodetic measurements is the indispensable foundation for all sustainable Earth observations, in situ as well as airborne and spaceborne, and underpins all georeferenced data used by society. The TRF is therefore of fundamental importance to geodesy in particular, science in general, and society as a whole.

The global network of GNSS receivers is essential to determining the TRF. Of the different space-geodetic

techniques, the GNSS network of observing stations is the densest. By colocating GNSS receivers with the stations of the other techniques it helps to integrate the separate technique-specific networks into one, integrated global observing system. GNSS also provides the means to access the TRF, allowing the absolute positions of GNSS receiver-equipped objects to be precisely given. Providing this ability to precisely position and navigate objects is one of the most important benefits of GNSS to science and society.

36.1.1 The International Association of Geodesy

The International Association of Geodesy (IAG), a founding association of the International Union of Geodesy and Geophysics (IUGG), is the international scientific organization devoted to the advancement of geodesy [36.5]. Its origin dates to 1862 when the Prussian General Johann Jacob Baeyer formed the central European arc measurement project with the ultimate goal of precisely determining the size and shape of the Earth. Today, more than 150 y later, the IAG continues to pursue this goal by advancing geodetic theory through research and teaching, by collecting, analyzing, modeling and interpreting observational data, by stimulating technological development, and by providing a consistent representation of the shape, rotation, and gravity of the Earth and planets including their temporal variations.

The IAG accomplishes its mission through the activities of its operating components, including its commissions, intercommission committees, services, and the global geodetic observing system (GGOS). Commissions represent the major fields of activity in geodesy and represent the IAG in all relevant scientific matters, promoting the advancement of science, technology, and international cooperation in these fields. The four IAG commissions are:

1. Reference frames
2. Gravity field
3. Earth rotation and geodynamics
4. Positioning and applications.

Intercommission committees address scientific matters that involve all of the commissions. There is currently one intercommission committee, the intercommission committee on theory.

Services organize the collection and reduction of geodetic observations and generate the geodetic products needed for scientific research and societal applications. The 14 IAG services span the relevant geometric,

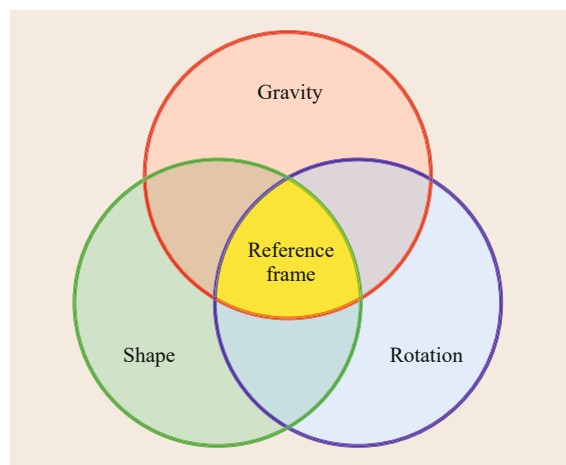


Fig. 36.1 Reference frames are determined from observations of the Earth's rotation, gravity, and shape. Reference frames also provide the means to integrate the three pillars of geodesy (rotation, gravity, and shape). These pillars are not independent of each other but are connected to each other by the common geophysical processes causing them to change. But to relate changes in the individual pillars to each other the changes must be given in the same reference frame (after [36.3, 4])

gravimetric, oceanographic, and related properties of the Earth. The geometric services of the IAG are the:

- International GNSS Service (IGS) (Chap. 33)
- International VLBI Service for Geodesy and Astrometry (IVS)
- International Laser Ranging Service (ILRS)
- International DORIS Service (IDS)
- International Earth Rotation and Reference Systems Service (IERS).

The gravimetric services of the IAG are the:

- International Gravity Field Service (IGFS)
- International Geoid Service (IGeS)
- International Gravimetric Bureau (BGI)
- International Center for Earth Tides (ICET)
- International Center for Global Earth Models (ICGEM)
- International Digital Elevation Model Service (IDEMS, to be confirmed).

The oceanographic services of the IAG are the:

- Permanent Service for Mean Sea Level (PSMSL)
- International Altimetry Service (IAS, to be confirmed).

The final service of the IAG, concerned with providing reference timescales, is the:

- Time Department of the International Bureau of Weights and Measures (BIPM).

36.1.2 The Global Geodetic Observing System

Recognizing the increasingly important role that geodesy plays in scientific research and societal applications, IAG established the global geodetic observing system (GGOS) in 2003, first as a project and then, in 2007, as a full component of the IAG. GGOS is meant to be *the* observing system of the IAG, organizing its technique-specific services under one unifying umbrella, thereby forming a comprehensive geodetic observing instrument integrating the hitherto separate pillars of geodesy (shape, rotation, and gravity) into one consistent observing system [36.6]. GGOS works with the other IAG components to provide unique, mutually consistent, and easily accessible geodetic constants, data and products for science and society. In addition, GGOS represents the IAG in the Group on Earth Observations (GEO) [36.7] and is IAG's contribution to the

Global Earth Observation System of Systems (GEOSS) that is being constructed by GEO.

GGOS provides the basis on which future advances in geosciences can be built. By considering the Earth system as a whole (including the geosphere, hydrosphere, cryosphere, atmosphere and biosphere), monitoring Earth system components and their interactions by geodetic techniques and studying them from the geodetic point of view, the geodetic community provides the global geosciences community with a powerful tool consisting mainly of high-quality services, standards and references, and theoretical and observational innovations. The mission of GGOS is [36.5]:

1. *To provide the observations needed to monitor, map and understand changes in the Earth's shape, rotation and mass distribution*
2. *To provide the global frame of reference that is the fundamental backbone for measuring and consistently interpreting key global change processes and for many other scientific and societal applications*
3. *To benefit science and society by providing the foundation upon which advances in Earth and planetary system science and applications are built.*

The goals of GGOS are [36.5]:

1. *To be the primary source for all global geodetic information and expertise serving society and Earth system science*
2. *To actively promote, sustain, improve, and evolve the global geodetic infrastructure needed to meet Earth science and societal requirements*
3. *To coordinate the international geodetic services that are the main source of key parameters needed to realize a stable global frame of reference and to observe and study changes in the dynamic Earth system*
4. *To communicate and advocate the benefits of GGOS to user communities, policy makers, funding organizations, and society.*

In order to accomplish its mission and goals, GGOS depends upon the services, commissions, and intercommission committees of the IAG. The services provide the infrastructure, data and products on which all contributions of GGOS are based. The commissions and intercommission committees provide expertise and support for scientific development within GGOS. In summary, GGOS is IAG's central interface to the scientific community and to society in general.

Organizational Structure

The components of GGOS are shown in Fig. 36.2. The governing components of GGOS are its consortium, coordinating board, and executive committee. These components serve as its steering committee, setting the strategic direction for GGOS. The coordinating office, like central bureaus of IAG services, oversees and coordinates the day-to-day activities of GGOS. It serves as the secretariat of GGOS and manages GGOS web services and outreach activities. The science panel is an independent, multidisciplinary advisory board. It provides scientific advice and support to GGOS to ensure that GGOS remains focused on relevant scientific and societal needs. The GGOS interagency committee (GIAC) is a forum for coordinating and supporting the development, implementation, and operation of the geodetic infrastructure that is owned by governmental institutions. Membership in GIAC is open to any governmental organization that contributes resources to the operation and development of space-geodetic observing systems.

The Bureaus of GGOS

Along with the science panel and GIAC, the operating arms of GGOS are its bureaus and focus areas. GGOS

currently has two bureaus: (1) The Bureau of Networks and Operations (2) The Bureau of Products and Standards.

Bureau of Networks and Observations. The goal of the Bureau of Networks and Operations (BNO) is to pursue the implementation of a network of space-geodetic observing systems of sufficient global distribution and capability that it will meet the needs of science and society as identified by GGOS [36.6]. To achieve this goal the Bureau works closely with the IAG services. In fact, the bureau is a consortium of service representatives supported by working groups, of which there are three:

- 1. *Working group on satellite missions* keeps GGOS informed about relevant satellite missions and supports GGOS in advocating for new missions that are needed to meet its goals.
- 2. *Working group on data and information systems* promotes the use of metadata standards and conventions for geodetic data and advocates for the interoperability of geodetic data centers.
- 3. *Working group on performance simulations and architectural trade-offs* uses simulations to assess the

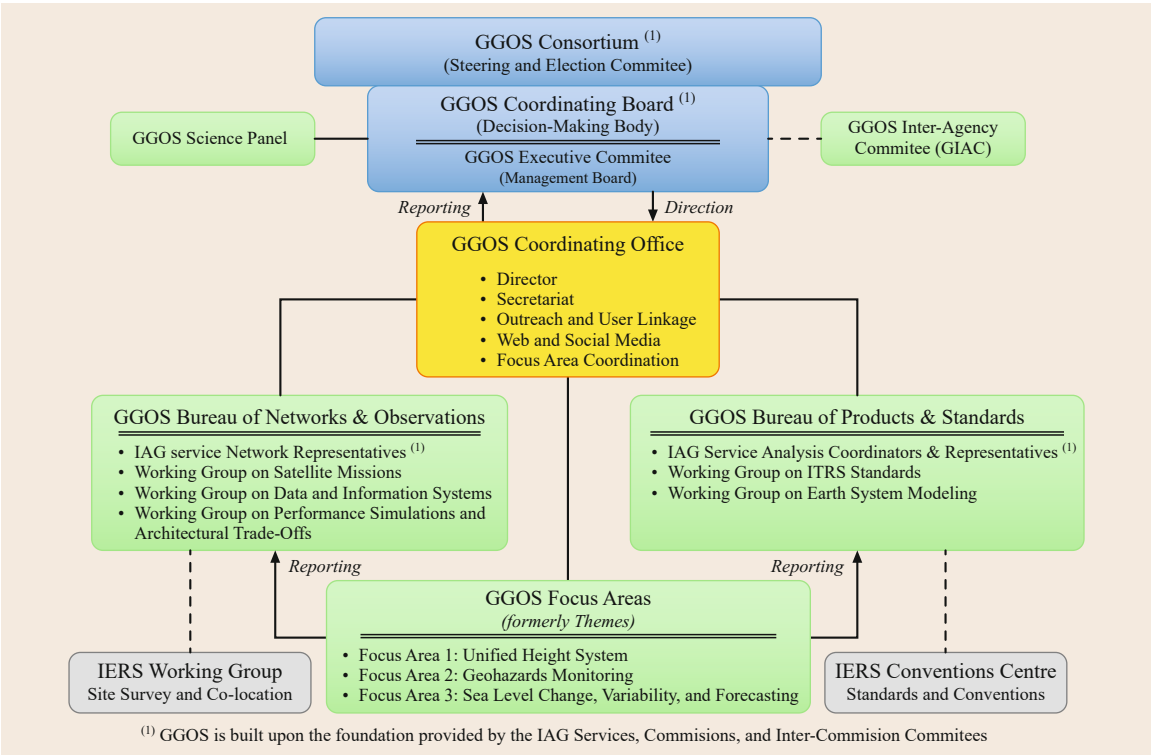


Fig. 36.2 The organizational structure of GGOS showing its governing components in blue, coordinating component in yellow, operating components in green, and affiliated components in gray (Courtesy of IAG-GGOS, reproduced under the CC BY-ND 4.0 license)

impact on geodetic data and products of different ground station architectures and their evolution, different space-based architectures and their evolution, and trade-offs between the ground- and space-based architectures including requirements on ground and space ties.

In pursuit of its goal, the BNO also works with the IERS working group on site survey and colocation. This working group is striving to improve the accuracy of the measurements of the relative positions of the reference points of colocated space-geodetic stations.

Bureau of Products and Standards. The goal of the Bureau of Products and Standards (BPS) is to make sure that the same standards and conventions are used by all components of the IAG. When combining data and products from different analysis centers or from different observing systems it is critical that they be determined using the same standards and conventions. Otherwise, inconsistencies can be introduced that can limit the accuracy of the combined data and products. As a first step towards meeting its goal, the BPS is compiling an inventory of the standards and conventions used by the organizations that generate IAG data and products. To help achieve its goal, the BPS has two working groups:

1. *Working group on ITRS standards* is pursuing the establishment of a new ISO (International Organization for Standardization) standard on global geodetic reference systems like the ITRS.
2. *Working group on Earth system modeling* is developing an integrated Earth system model that will apply to all observation techniques and all pillars of geodesy (rotation, gravity, and shape).

In pursuit of its goal, the BPS works closely with the IERS conventions center, that component of the IERS that is responsible for maintaining the constants, standards, and conventional models used by the IERS.

The Focus Areas of GGOS

GGOS focus areas are interdisciplinary in nature and address broad and critical issues that are important to science and society and that geodesy can contribute to but that need further development by or coordination within the geodetic community. GGOS currently has three focus areas:

1. Unified height system
2. Geohazards monitoring
3. Sea level change, variability and forecasting.

Unified Height System. The goal of a number of IAG working groups during the last few decades has been the unification of the more than 100 existing vertical reference systems. This involves defining and realizing a global reference level and determining the transformations between local height datums and the global, unified one. When this is achieved, all physical heights will be referred to the same global reference level. To aid in the realization of this goal, GGOS created a focus area, focus area 1, on the unified height system. To date, the activities of focus area 1 have been focused on determining a reliable value for the reference geopotential W_0 that can be used for the conventional reference level when realizing a global height system.

Geohazards Monitoring. Helping to mitigate the impact on human life and property of natural hazards such as earthquakes, volcanic eruptions, debris flows, landslides, land subsidence, tsunamis, floods, storm surges, hurricanes and extreme weather is one of the most important services that geodesy can provide to science and society. Since natural hazards often cause objects to be displaced and the Earth's surface to be deformed, GNSS plays a crucial role in this. For example, GNSS can be used to monitor the pre-eruptive deformation of volcanoes and the preseismic deformation of earthquake fault zones, aiding in the issuance of volcanic eruption and earthquake warnings. GNSS can also be used to rapidly estimate earthquake fault motion, aiding in the modeling of tsunami genesis and the issuance of tsunami warnings. GNSS observations are essential for understanding the processes causing the hazard, for assessing the risks of the hazard, for monitoring the development of the hazard, for deciding whether or not to issue an early warning, and to support rescue and damage assessment activities.

Recognizing the important role that geodetic observations play in disaster prevention and mitigation, GGOS created a focus area, focus area 2, on geohazards monitoring. The objective of focus area 2 is to improve the effectiveness of the geodetic community in supporting natural hazard identification, assessment, prioritization, prediction, and early warning. As an international organization, GGOS can be a very effective advocate for the role of geodesy in understanding and mitigating natural hazards. GGOS can also be an effective advocate for improving the geodetic data needed for natural hazards research including better spatial coverage, higher sampling rate, lower latency, and wider data availability, particularly of synthetic aperture radar (SAR) and GNSS data.

Sea Level Change, Variability and Forecasting. In 1990, 23% of the world's population lived both less

than 100 km from the coast and less than 100 m above sea level. Nearly a quarter of the world's population is therefore vulnerable to the effects of a rising sea level. Since the long-term average rate of sea level rise is only a few mm/y, mitigation efforts can be planned well in advance. But great demands are placed on geodetic observing systems because the sea level rise signal is so small. For example, the terrestrial reference frame, which should be at least an order of magnitude more accurate than the amplitude of the signal being measured, needs to be accurate and stable to within about 0.1 mm/y to support studies of sea level change. This makes sea level change studies one of the most demanding applications of geodetic observing systems.

Recognizing the important role that geodetic observations play in sea level change studies, GGOS created a focus area, focus area 3, on sea level change, variability, and forecasting. The objective of focus area 3 is to improve our understanding of the causes and consequences of sea level change through the application of geodetic measurements.

GGOS and Reference Frames

As discussed above, GGOS is built upon the foundation provided by the IAG services, commissions, and inter-commission committees. The IAG services coordinate the acquisition and analysis of geodetic observations of the Earth's time varying gravity, rotation, and shape (Fig. 36.1). An important goal of GGOS is to advocate for the improvement of the global geodetic infrastructure, including the GNSS infrastructure, that provides the geodetic observations. One of the most scientifically and societally important applications of geodetic observations is their use for determining reference frames. This chapter of the GNSS Handbook discusses the use of geodetic observations, especially GNSS observations, to determine reference frames and to study changes in the rotation of the Earth. The use of geodetic observations to study changes in the shape of the Earth are discussed in the chapters in the Handbook concerned with precise positioning (Chap. 25), generation of orbit products (Chap. 34) and geodynamics (Chap. 37).

36.2 Global and Regional Reference Frames

This section is divided in four parts. The first part introduces the types of reference frame representations for a deformable Earth, taking into account all sorts of linear and nonlinear motions. The second part deals with global reference frames, focusing on the International Terrestrial Reference Frame (ITRF), its derivatives formed by the International GNSS Service (IGS) and the IGS contribution to the ITRF construction, describing the fundamental role of the IGS network in connecting the three other techniques: very long baseline interferometry (VLBI), satellite laser ranging (SLR) and Doppler orbitography radiopositioning integrated by satellite (DORIS). The third part details the GNSS-based global and regional reference frames and how these frames are linked to the global ITRF, through the usage of IGS products. The fourth part gives general guidelines on how to realize GNSS-based local, regional and global reference frames, fully consistent with and optimally aligned to the ITRF.

36.2.1 Reference Frame Representations for the Deformable Earth

The Earth is a complex dynamic system that undergoes deformations caused by various geophysical processes that should be taken into account when constructing a reference frame. The frame is implemented through a geodetic network anchored to the Earth's crust and

therefore can be called crust-based frame. The expression of the instantaneous station position $X(t)$, at epoch t , can be written as the sum of its regularized position $X_R(t)$ and high frequency geophysical variations $\Delta X_i(t)$ (see IERS conventions [36.8], Chap. 4)

$$X(t) = X_R(t) + \sum_i \Delta X_i(t), \quad (36.1)$$

and

$$X_R(t) = X_R(t_0) + \dot{X}_R(t - t_0), \quad (36.2)$$

where t_0 is the reference epoch of the station position and \dot{X}_R its linear velocity. $X_R(t)$ is introduced here in order to obtain a position with more regular time variation, after removing high-frequency time variations caused by geophysical processes using conventional corrections $\Delta X_i(t)$.

Chapter 7 of the IERS conventions [36.8] provides a full description of the currently agreed-upon conventional models that go into $\Delta X_i(t)$, such as Earth tide, ocean loading, atmospheric pressure, and so on, and used by the analysis centers (ACs) dealing with space geodesy data. In addition to these conventional recommended models, other geophysical phenomena have a large impact on space geodesy observations and therefore need to be taken into account in reference frame

implementation. We can categorize the resulting deformations of such phenomena at the surface of the deformable Earth into the following two types:

- Nearly linear motions that can be expressed in the mathematical geodesy formulation as constant with time. They are caused by two main types of processes: plate tectonics and glacial isostatic adjustment (GIA). Plate tectonics induce mainly horizontal motion, which is traditionally modeled via a rotation pole for each plate involving horizontal velocity components [36.9], while GIA implies horizontal and vertical deformations.
- Nonlinear motions that include periodic signals (e.g., annual, semi-annual or interannual) that are caused by nontidal loading effects due to the atmosphere, ocean circulation, terrestrial hydrology and ice melting [36.9]; ruptures provoked by earthquakes and volcanic eruptions; and slow transients or postseismic deformations.

Taking into account the above two types of motions, two categories of reference frame representations can be introduced: quasi-instantaneous reference frame and long-term or secular reference frame.

Quasi-Instantaneous Reference Frames

A quasi-instantaneous frame gives access to average station positions, using short timespan of space geodesy observations: commonly one day, and up to one week. In this case station positions are only valid at the central epoch of the observations used. More than one week of observations could of course mathematically be used, but the resulting averaged station positions would then be biased by tectonic motion effects. Long time series of quasi-instantaneous frame solutions naturally contain all types of linear and nonlinear station motions. The analysis and accumulation (rigorous stacking) of long time series of quasi-instantaneous frame solutions permit not only the study of all types of linear and nonlinear station motions they naturally contain, but also the construction of a long-term secular frame, such as the ITRF.

Examples of quasi-instantaneous reference frames are daily or weekly solutions provided by the analysis and combination centers of the IAG services of the four space geodesy techniques. In the particular case of GNSS, in addition to IGS global solutions, local, national and regional solutions are also produced by research groups for scientific studies and by institutions in charge of the maintenance of national reference frames based on GNSS permanent networks. In general, IGS analysis center daily or weekly solutions are generated by estimating not only station positions and

EOPs, but also orbits, clocks and eventually other parameters such as troposphere gradients. Local, national and regional solutions are generally computed by fixing IGS products (orbits, clocks and EOPs) where the main target is the estimation of station positions, using either a network approach (all stations are adjusted together) or precise point positioning approach, on a station-by-station basis. General guidelines are provided in Sect. 36.2.4 on how to express or align a GNSS solution into the ITRF.

Long-Term Secular Reference Frames

A long-term or secular frame gives access to station positions at a given epoch t_0 and station linear velocities. Examples of long-term reference frames are the ITRF (Sect. 36.2.2) and a cumulative solution obtained by stacking time series of quasi-instantaneous reference frames. The choice of t_0 does not mathematically matter, but should be selected to be close to the central epoch of the stacked time series. The users can actually propagate station positions and their associated variances from the reference epoch t_0 to any other epoch t . For a given station with position vector $\mathbf{X}(t_0)$ at epoch t_0 and velocity vector $\dot{\mathbf{X}}$, its position $\mathbf{X}(t)$, at epoch t is given by

$$\mathbf{X}(t) = \mathbf{X}(t_0) + \dot{\mathbf{X}}(t - t_0), \quad (36.3)$$

and the variance propagation law gives its variance at epoch t as

$$\begin{aligned} \text{var}(\mathbf{X}(t)) = & \text{var}(\mathbf{X}(t_0)) + 2(t - t_0) \text{cov}(\mathbf{X}, \dot{\mathbf{X}}) \\ & + (t - t_0)^2 \text{var}(\dot{\mathbf{X}}). \end{aligned} \quad (36.4)$$

The stacking of time series of quasi-instantaneous reference frame solutions is usually operated using (36.3) – a type of equation where the unknowns are station positions $\mathbf{X}(t_0)$, and station velocities $\dot{\mathbf{X}}$. Equation (36.3) could also be generalized to include transformation parameters in order to account for possible reference frame differences between individual quasi-instantaneous reference frames themselves (which might not have the same origin, scale and/or orientation) and with respect to the stacked/combined long-term solution. Minimum constraints equations as described in Sect. 36.2.4 could also be added to the stacking model ((36.3)-type) in order to express a cumulative GNSS long-term solution in the ITRF.

Geocenter Motion and Periodic Signals

A GNSS satellite, as any satellite, is theoretically orbiting around the center of gravity, or the center of mass (CM) of the total Earth system. Therefore in

theory, the instantaneous CM position reflects the natural origin of the inertial frame in which the satellite orbit is expressed. Analysis of satellite geodesy data have clearly indicated for about two decades that the network of stations attached to the Earth's crust and materializing the reference frame has detectable translational motion with respect to CM, known as the geocenter motion [36.10]. This motion is often defined as the motion of the CM with respect to the center of figure (CF) of the solid Earth surface [36.11] and is believed to be the crust response to various geophysical fluid displacements within the Earth system, such as the atmosphere, oceans, terrestrial hydrology and ice sheets. It is assumed to include tidal, nontidal and secular components. The tidal parts of the geocenter motion induced by the atmosphere and oceans, with an amplitude that may reach up to 1 cm, are included in the models recommended by the IERS conventions [36.8] to be taken into account a priori in the station displacement of space geodesy techniques. The nontidal part of the geocenter motion, with an amplitude of a few mm, manifests itself in the form of periodic signals: annual, semi- and interannual, and is quantified through data analysis of time series of station positions determined by space geodesy techniques, or through external geophysical models. The secular part, often called the geocenter velocity, is believed to be less than 1 mm/y [36.12]. The detailed review by [36.12] is the most extensive article describing the theory of the geocenter motion and its geophysical implications, as well as its quantification over different timescales.

There are basically three main methods for estimating the nonlinear geocenter motion components:

1. The translational
2. The degree-1 load-induced deformation
3. The inverse approaches.

The translational approach consists in estimating the three equatorial components of the CF, which is in fact approximated by the barycenter of the implied geodetic network, often called the center of network (CN), with respect to CM. The translational approach is called a kinematic approach when the degree-1 coefficients of the gravity field are estimated, which are proportional to the geocenter motion components [36.13]. It is also called a network shift approach when the seven- or six-parameter similarity (Helmert) transformation formula (see for instance (36.7)) is used to infer the three translation components between a time series of quasi-instantaneous frames and a secular long-term frame such as the ITRF. Indeed, the ITRF origin is

defined by the long-term average of the SLR CM realization. Fitting a sine and/or a cosine function yields in fact the amplitude and phase of annual and/or semi-annual signals of the geocenter motion.

Although the SLR technique suffers from its poor spatiotemporal network, leading to the so-called *network effect* when using the translational approach [36.14], it is the most precise space geodetic technique for the geocenter nonlinear motion estimation.

The estimability of the geocenter motion by GNSS via the kinematic or network shift approaches faces intrinsic complications due to an inherent coupling of the GNSS orbit dynamic parameters; [36.15, 16] showed that the GNSS geocenter Z-component is strongly correlated to a particular parameter of the solar radiation pressure. In [36.17] it was demonstrated, via a collinearity diagnosis formalism, that the inability of GNSS, as opposed to SLR, to properly sense the location of the geocenter CM is mostly explained by the estimation, in the GNSS case, of epoch-wise station and satellite clock offsets simultaneously with tropospheric parameters.

The degree-1 approach was first introduced in [36.18], using the spherical harmonics formalism, to infer not only the translational geocenter motion, but also the accompanying load-induced crust deformation using GNSS (GPS) time series of quasi-instantaneous frames of a globally distributed network of stations. They in fact demonstrated that the translational components of the geocenter motion are functions of degree-1 coefficients of the associated load deformation. It was then shown in [36.19] that the truncated higher-degree terms of the harmonic expansion of the load-induced deformation alias significantly into the degree-1 terms, and therefore higher-degree terms, up to 50, must be included in the estimation, leading so to the so-called inverse approach. Many authors (as referenced in [36.11]) have expanded and improved the inverse approach and its application, using not only GNSS(GPS) data that suffer from the network sparseness in ocean areas, but also data-assimilated ocean bottom pressure (OBP) models as well as GRACE gravity data.

In addition to geophysical (load-induced) periodic signals that explain about half of the observed GNSS seasonal power, other signals are also frequent and detected in the GNSS residuals of station position time series, such as the GPS draconitic errors [36.20]. The GPS draconitic year (of 351.2 d), is the period for the GPS orbit constellation to repeat its orientation with respect to the Sun. Harmonic signals of this period have actually been observed in the power spectra of nearly all IGS products [36.21].

36.2.2 Global Terrestrial Reference Frames

Following the terminology adopted by the geodetic community since the advent of space geodesy, we distinguish between a terrestrial reference system (TRS) and a terrestrial reference frame (TRF). While the former has a mathematical and physical foundations for its definition and properties, the latter represents its numerical realization constructed upon space geodesy observations (hence with uncertainties) and is accessible to the users through numerical values (e.g., positions as a function of time of a network of Earth crust-based points). The main physical and mathematical properties of a TRS (at the theoretical level) or of a TRF (at the realization level) are the origin, the scale, the orientation and their time evolution. The latter is usually expressed through rates (time variations) of translations (origin components), scale and rotations (orientation parameters).

While the origin and the scale (having physical properties) are the most critical parameters of interest to Earth science applications, the orientation and its time variation are of least consequence because they are arbitrarily and conventionally defined. In fact adopting a given orientation of the three axes of the reference system is a matter of conventions and convenience and would not change the relative shape of the implied geodetic network used to create the reference system. Continuous and long-term space geodesy observations are crucial for realizing a TRS that is able to precisely characterize and model Earth surface movements, such as tectonic plate motion. In the absence of technique-specific systematic errors, and if all geophysical processes are accurately accounted for in geodetic analysis, TRF origin and scale should be stable over time, i.e., should not exhibit any drift or discontinuities over the entire timespan of the implied geodetic observations.

None of the space geodesy techniques is able to provide all the necessary parameters for the TRF definition (origin, scale and orientation). While satellite techniques are sensitive to the Earth center of mass (a natural TRF origin; the point around which a satellite orbits), VLBI (whose TRF origin is arbitrarily defined through some mathematical constraints) is not. The scale is dependent on the modeling of some physical parameters, and the absolute TRF orientation (unobservable by any technique) is arbitrarily or conventionally defined through specific constraints. The utility of multitechnique combinations is therefore recognized for reference frame determination, and in particular for accurate reference definition. In principle, the particular strengths of one observing method can compensate for weaknesses in others if the combination is properly con-

structed, suitable weights are found, and accurate local ties in colocation sites are available.

The key element of a multitechnique combined frame, as used for the ITRF, is the availability of a sufficient number of globally distributed colocation sites. A colocation site is defined by the requirement that two or more space geodetic distinct instruments are operating at the same location or at locations very close to one another, which are very precisely surveyed in three dimensions, using geodetic classical surveys or the GPS technique. Classical surveys are usually direction angles, distances, and spirit leveling measurements between instrument reference points or geodetic markers. Adjustments by least squares of local surveys are generally performed by national geodetic agencies operating space geodesy instruments, yielding differential coordinates (local ties) connecting the colocated instrument reference points.

Figure 36.3 shows the four-technique colocation site at Yarragadee (western Australia), with the modern 12 m VLBI radio-telescope that started its operation in 2011, the National Aeronautics and Space Administration (NASA) SLR MOBLAS 5 system, the DORIS beacon, the GNSS pillars (called YARR, YAR2, YAR3) and the Gravimeter hut for campaign gravimetry measurements.

Intermarker distance and accuracy of the local tie are the two main criteria that must be considered for the definition of a colocation site [36.22]. Given the need for local tie vectors to be precise at the 1 mm level, and considering the increase in atmospheric refraction as a function of increased station separation, the distances between geodetic markers at colocation sites should not exceed 1 km. In addition, repeated surveys of the marker footprint are necessary for long-term local tie stability. The typical uncertainty of the local ties used for the ITRF is 2–5 mm (sometimes larger than 5 mm for the less precise ties). From the ITRF experience,

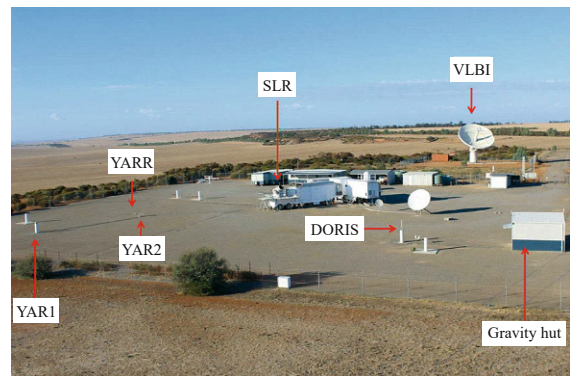


Fig. 36.3 Yarragadee (western Australia) four-technique colocation site (courtesy of Geoscience Australia)

discrepancies between local ties and space geodesy estimates are frequent as discussed in the following section. However, discrepancies mean that either local ties or space geodesy estimates (or both) are imprecise or in error. One of the major local tie limitations is in fact to precisely determine the eccentricity between the external physical reference point used by the surveyors and the point referenced by space geodesy data analysts, for example the intersection of axes of VLBI or SLR telescopes, the DORIS beacon or the electrical GNSS antenna phase center [36.23]. The estimated uncertainty for each internal instrument offset is probably not better than 2 mm, and consequently, the overall local tie error would be at best 3 mm per component.

The International Terrestrial Reference System (ITRS) and Frame (ITRF)

The international terrestrial reference system (ITRS) was developed by the geodetic community for the most demanding scientific applications under the auspices of the IERS. Following the IERS conventions and its updates [36.8, Ch. 4], the ITRS definition fulfills the following conditions:

1. It is geocentric, its origin being the center of mass for the whole Earth, including oceans and atmosphere.
2. The unit of length is the meter (SI). The scale is consistent with the geocentric coordinate time (TCG) for a geocentric local frame, in agreement with International Astronomical Union (IAU) and IUGG (1991) resolutions. This is obtained by appropriate relativistic modeling.
3. Its orientation was initially given by the Bureau International de l'Heure (BIH) orientation at 1984.0.
4. The time evolution of the orientation is ensured by using a no-net-rotation condition with regards to horizontal tectonic motions over the whole Earth.

The most accurate realizations of the ITRS is called the international terrestrial reference frame (ITRF). The implementation of the ITRF is fundamentally based on the rigorous combination of geodetic products of the main space geodetic techniques (GNSS, VLBI, SLR, and DORIS), through their colocated measuring instruments at a certain number of core sites. The ITRF combination model is based on the linearized form of the general similarity transformation formula, as it will be detailed below.

There is no single ITRF, but rather a series of updated and improved versions of ITRF. The versions are identified by the year associated with the date of last data used in the analysis, and should not be con-

fused with the date of applicability. The most recent versions are ITRF97, ITRF2000, ITRF2005 and the ITRF2008 [36.24–26]. Generally, as time progresses, there is less need for frequent updates, because more time may be needed to make significant improvements through the addition of new data and improved models. However, to satisfy increasing accuracy requirements, the ITRF will continue to be updated to incorporate more advanced models for the time-dependent reference coordinates. Since the tracking network equipped with the instruments of those techniques is evolving and the period of data available increases with time, the ITRF is constantly being updated.

For more than ten years, initiated first by the IGS, analysis centers of the three other space geodesy techniques (VLBI, SLR, DORIS) have made available time series of station positions and Earth orientation parameters (EOPs) in SINEX (software independent exchange) format [36.27]. The power of times series of station positions, allowing the control not only of the station behavior and in particular to monitor nonlinear motion, but also the frame physical parameters (origin and scale), led the ITRF center to consider them as input for the ITRF generation, starting with the ITRF2005 [36.25]. In addition to station positions and velocities, ITRF2005 and ITRF2008 [36.26] integrate also consistent daily EOPs. The latter was already used by the IERS EOP center in order to improve the consistency of the IERS operational series of EOPs with the ITRF [36.28]. Up to the ITRF2008, the ITRF input time series solutions are provided on a weekly basis by the IAG international services of satellite techniques: the IGS [36.29], the ILRS [36.30] and the IDS [36.31], and on a daily (VLBI session-wise) basis by the IVS [36.32]. Each per-technique time series is already a combination of the individual analysis center solutions of that technique. As an example, the GNSS (mainly GPS) submitted solution to the ITRF2008 is a combination of the first reprocessed solutions by the IGS analysis centers and covers the time period 1997.0–2009.5 [36.33]. Note that a very small portion of Global'naya Navigatsionnaya Sputnikova Sistema (GLONASS) observations were used by some IGS ACs that contributed to the reprocessing effort. Starting on 19 August 2012, the IGS switched to daily integration and therefore daily IGS SINEX files will be used in the future ITRF solutions.

The procedure adopted for the ITRF formation involves two steps [36.25, 26, 34]:

1. Stacking the individual time series to estimate a long-term solution per technique comprising station positions at a reference epoch, station velocities and daily EOPs

- Combining the resulting long-term solutions of the four techniques together with the local ties in colocation sites.

The main two equations of the combination model are given below. They involve a 14-parameter similarity transformation, station positions and velocities, and EOPs and are written as

$$\begin{cases} \mathbf{X}_s^i = \mathbf{X}_c^i + (t_s^i - t_0) \dot{\mathbf{X}}_c^i \\ \quad + \mathbf{T}_k + D_k \mathbf{X}_c^i + \mathbf{R}_k \mathbf{X}_c^i \\ \quad + (t_s^i - t_k) [\dot{\mathbf{T}}_k + \dot{D}_k \mathbf{X}_c^i + \dot{\mathbf{R}}_k \mathbf{X}_c^i] \\ \dot{\mathbf{X}}_s^i = \dot{\mathbf{X}}_c^i + \dot{\mathbf{T}}_k + \dot{D}_k \mathbf{X}_c^i + \dot{\mathbf{R}}_k \mathbf{X}_c^i, \end{cases} \quad (36.5)$$

$$\begin{cases} x_s^p = x_c^p + R_{yk} \\ y_s^p = y_c^p + R_{xk} \\ UT_s = UT_c - \frac{1}{f} R_{zk} \\ \dot{x}_s^p = \dot{x}_c^p \\ \dot{y}_s^p = \dot{y}_c^p \\ LOD_s = LOD_c, \end{cases} \quad (36.6)$$

where for each point i , \mathbf{X}_s^i (at epoch t_s^i) and $\dot{\mathbf{X}}_s^i$ are positions and velocities of technique solution s and \mathbf{X}_c^i (at epoch t_0) and $\dot{\mathbf{X}}_c^i$ are those of the combined solution c . For each individual frame k , as implicitly defined by solution s , D_k is the scale factor, \mathbf{T}_k the translation vector and \mathbf{R}_k the rotation matrix. The dotted parameters designate their derivatives with respect to time. The translation vector \mathbf{T}_k is composed of three origin components, namely T_x , T_y , T_z , and the rotation matrix of three small rotation parameters: R_x , R_y , R_z , following the three axes, respectively x , y , z . t_k is a conventionally selected epoch of the seven transformation parameters.

In addition to (36.5) involving station positions (and velocities), the EOPs are added by (36.6), making use of pole coordinates x_s^p , y_s^p and universal time UT_s as well as their daily rates \dot{x}_s^p , \dot{y}_s^p and length-of-day LOD_s , where $f = 1.002737909350795$ is the conversion factor from universal time (UT) into sidereal time. The link between the combined frame and the EOPs is ensured via the three rotation parameters appearing in the first three lines of (36.6).

Note that (36.5) uses the linearized form of the general similarity transformation formula, neglecting second- and higher-order terms [36.8, 35].

In the first step of the ITRF construction, the first two lines of (36.5) and the entire (36.6) are used to estimate long-term solutions for each technique, by accumulating (rigorously stacking) the individual technique time series of station positions and EOPs. In the second step, the entire two equations are used to combine the long-term solutions obtained in step 1, together with local ties in colocation sites.

The number of colocation sites has evolved with time since the start of the ITRF combination activities in 1984, due to the decommission of certain historical sites for multiple reasons, and the appearance of a few new colocation sites. Figure 36.4 illustrates the distribution of the total number of VLBI, SLR and DORIS operating sites in 2015, as well as the IGS/GNSS collocated sites: IVS is managing a network of 49 radio telescopes located in 46 sites, ILRS 38 laser telescopes in 37 sites, IDS 53 beacons in 53 sites, and IGS more than 400 GNSS permanent, continuously operating receivers/antennas. All in all there are 90 collocated sites: 30 GNSS-SLR, 38 GNSS-VLBI and 43 GNSS-DORIS collocations. There are only 11 sites where VLBI and SLR are collocated, nine in the northern and only two in

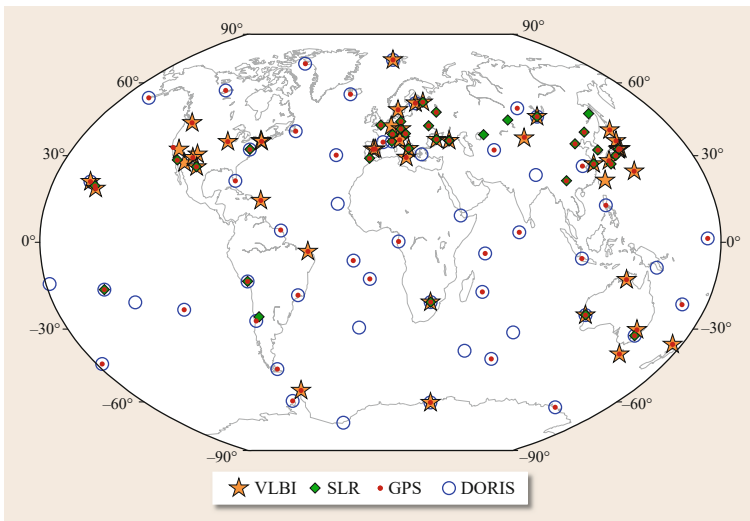


Fig. 36.4 VLBI, SLR and DORIS sites and their colocations with GPS

the southern hemisphere. Unfortunately more than half of the VLBI and SLR instruments are old generation systems. As a consequence, the improvement of the underlying geodetic infrastructure of ITRF is an important goal of GGOS [36.6], discussed in the previous section. The low number, the nonoptimal coverage, and the low performance of some of the 11 VLBI-SLR colocation sites are significant limiting factors to ensuring a precise connection between these two techniques in the ITRF implementation. In fact, GNSS is playing a major role in connecting the three other techniques, given the fact that almost all SLR and VLBI sites, as well as 43 DORIS sites are colocated with permanent IGS stations. The drawback of this situation is that if there is any GNSS-related bias, this will contaminate the parameters that define the ITRF, primarily the origin and the scale that are determined by SLR and VLBI. There most probably are other technique-specific errors related to the mismodeling of the instrumental measurement reference points, not only for GPS [36.23], but also for the other techniques. Indeed, based on ITRF2008 results [36.26], tie discrepancies for 47, 43 and 34% of the total local tie vectors between GPS-VLBI, GPS-SLR and GPS-DORIS respectively are larger than 6 mm, corresponding to the level of scale agreement between VLBI and SLR solutions included in the ITRF2008 adjustment.

IGS Reference Frames and Their Relationship with the ITRF

The IGS products were integrated in the IERS combined products in 1992 and have contributed since then to the ITRF starting with ITRF91 [36.36]. All the IGS products are expressed in and are consistent with the ITRF frames. At the inception of its activities, the IGS used directly the ITRF frames to be the underlying frame of its products [36.37–39]. Following the methodology of [36.38, 40], the IGS started in 2000 to form its own, internally more consistent GPS-only frame, but still inheriting the ITRF definition in terms of origin, scale and orientation [36.41]. A more detailed history of IGS reference frame realizations can be found in [36.42, 43]. Starting with GPS week 1400 (5 November 2006), the IGS switched from relative to absolute model corrections to account for antenna phase center variations (PCV) [36.44]. At the same time, the IGS adopted directly the ITRF2005 [36.25] to form its specific frame called IGS05, composed of about 100 sites whose coordinates were corrected to account for relative to absolute PCV differences. In order to preserve the ITRF2005 origin, scale and orientation, the IGS05 was aligned to the ITRF2005 using 14-parameter similarity transformation [36.33]. In reality, among the 14 parameters, only the scale factor

was significant, representing the mean of the height relative to absolute differences over the IGS05 stations. On 17 April 2011, the IGS generated and adopted the IGS08 frame [36.45], derived from ITRF2008. The IGS08 is composed of positions and velocities of a reference set of 232 stable GNSS stations extracted from ITRF2008, where corrections were applied to 65 stations to ITRF2008 positions in order to comply with the antenna calibration models used in present-day GNSS data analysis [36.23] (igs08.atx, in use since GPS week 1632). On 17 October 2012, the IGS updated the IGS08, called IGB08 [36.45], by adding about 36 stations in replacement of some decommissioned or dormant IGS08 stations. It should be noted that ITRF2008, IGS08 and IGB08 are however equivalent at the global level (sharing the same underlying origin, scale and orientation), although station-dependent position differences can exist.

36.2.3 GNSS-Based Reference Frames and Their Relationship with the ITRF

This section deals with reference frames that are built using GNSS data only, but are nominally aligned with the ITRF in origin, scale and orientation. The first subsection presents the GNSS-specific frames that are implemented by the different GNSS providers and in which the broadcast orbits are expressed. The second subsection discusses the regional reference frames that are also based on GNSS data only, while aligned to the ITRF via common processed stations. The third subsection develops general and mathematical guidelines on how to optimally align global or regional frames to the ITRF using IGS products.

GNSS-Specific Reference Frames

In order to ensure the integrity of any GNSS system and to precisely determine satellite orbits of its constellation, a specific reference frame has to be defined and maintained over time. The computed orbits are then transmitted to the users via the GNSS navigation message that allow determination of the user location, which will be expressed in the reference frame of that of the used orbits. The GNSS systems and frames in existence with publicly available information and publications are WGS84 for GPS, whose newest realization is designated as G1674 [36.46], PZ-90 for GLONASS whose latest realization is PS-90.11 [36.47] introduced in early 2014, CGCS2000 for COMPASS [36.48], the Galileo terrestrial reference frame (GTRF) for Galileo where the first series of its realization is described in [36.49], the newest one being designated as GTRF14v01, and the Japanese

geodetic system (JGS) for quasi-zenith satellite system (QZSS), which is believed to be consistent with or close to the newest Japanese geodetic datum 2011 (JGD2011) that was revised after the 2011 Tohoku Earthquake [36.50].

In an effort to ensure the interoperability of timing and geodetic references among the different GNSSs, working group D of the International Committee on GNSS (ICG) is actively interacting with the GNSS providers toward a more rigorous and accurate alignment of the GNSSs to a common time reference and to the ITRF. To our best knowledge, all recent and up-to-date realizations of all GNSS-specific geodetic reference systems are believed to be aligned to ITRF2008. However, almost all these realizations, except the GTRF series, are based on GNSS data with short timespans, most often a few days or one week of observations. While the GTRF series are aligned to the current ITRF version at the few millimeter level in both positions and velocities, the other GNSS-specific frames are obtained via the adjustment of station positions at the central epoch of the observations used, with no velocity estimates to account for time variations. Depending on the selected reference epoch of the adjusted positions of the control stations, the impact of, for example, tectonic motion will be at the few centimeter per year level. If we consider a scenario of 10 y before a new update of the control station positions is made, with no plate motion model applied, a 20–70 cm position error will be accumulated and mapped into the computed orbits, depending on the station locations. Consequently, we believe that the current realizations of GNSS-specific frames agree to each other and with the ITRF2008 at the few decimeter level. However, this level of agreement is certainly well below the inherent and typical uncertainty of the broadcast orbits. Over a ten-month period, [36.51] analyzed signal-in-space ranging errors (SISREs) for all current GNSS systems and showed that the global average SISRE values amount to 0.7 m (GPS), 1.5 m (BeiDou), 1.6 m (Galileo), 1.9 m (GLONASS), and 0.6 m (QZSS). As a consequence, the position of a real-time user, with single or multi-GNSS capabilities, is at the level of 1–2 m accuracy.

A way forward to improving the consistency of GNSS-specific frames at the few millimeter level is to follow the GTRF example, or to disclose data to the IGS of a subset of stations used in the ground segment, as it was the case of 11 stations of the National Geospatial Intelligence Agency (NGA) for the US Department of Defense, which were included in the ITRF2008 [36.26, 46]. The remaining challenge rests, however, in improving the intrinsic accuracy of the broadcast orbits of all GNSS constellations.

Regional and National Terrestrial Reference Frames

Since the start of the ITRF development, together with the advent of GNSS positioning performance, significant effort was and is still undertaken by national mapping agencies to redefine and modernize continental and national geodetic systems, so that they are compatible with the global ITRF.

The structure of IAG commission 1 (reference frames) includes a subcommission 1.3 dealing with the definitions and realizations of regional reference frames and their connection to the global ITRF. The commission offers a home for service-like activities addressing theoretical and technical key common issues of interest to regional organizations. Six regional organizations are part of IAG subcommission 1.3, distributed to cover all continents (AFREF for Africa, NAREF and SIRGAS for North and South Americas, EUREF for Europe, APREF for Asia and Pacific, and SCAR for Antarctica).

Regional reference systems and frames are defined with respect to the ITRS/ITRF, realized and maintained by the IAG regional entities; the best known and advanced ones are ETRS89 for Europe, NAD83 for North America, and SIRGAS for South America. These regional entities usually play a major role in redefining regional and national geodetic systems and their relationship to the ITRF. In addition, many countries have already redefined or are in the process of redefining their geodetic systems, directly connected to the ITRF, using their national permanent GNSS networks. The main purpose of regional and national reference frames is for georeferencing applications with centimeter precision and accuracy. There are three main categories of implementation of these reference frames:

1. Station positions at a given epoch, eventually updated more or less frequently. This is the case, for example, for NAD83 [36.52, 53] and SIRGAS [36.54] for North and South America respectively, and GDA94 for Australia [36.55]
2. Station positions and minimized velocities. This is the case of ETRS89 for Europe where velocities are minimized by removing the angular velocity of the Eurasian plate when transforming from ITRF to ETRS89 realization [36.56]
3. Station positions and deformation model corrections. A case example is the New Zealand geodetic datum 2000 (NZGD2000) [36.57] where a deformation model is elaborated to correct coordinates for the effect of regional-scale tectonic movements for all geodetic reference points. The accumulated displacements estimated by the deformation model allow the computation of station coordinates as if

they were observed at the fixed reference epoch of 2000.0.

36.2.4 General Guidelines for GNSS-Based Reference Frame Implementation

By design, the ITRF is to be regarded as a common global standard that provides the most accurate frame definition: long-term averages of the origin, scale and orientation, necessary for the consistency and interoperability of Earth science and societal applications. In the meantime, with the proliferation of dense GNSS networks at the local, national, continental and global levels, it is obviously impossible to include all worldwide permanent GNSS stations in the IGS (and consequently) in the ITRF networks. It becomes however desirable to express all local, national, continental and global GNSS network solutions in the ITRF.

In order to access the ITRF, it can be used directly, via its products (station positions and velocities), but also indirectly, using IGS products. In the following we describe general guidelines that allow the efficient expression of a GNSS-based solution of station positions in the ITRF, using IGS products (orbits, clocks and EOPs). This method, based on the equations of minimum constraints (MC) (see for instance [36.34, 58]), is described below for the case of an epoch solution (as a materialization of a quasi-instantaneous reference frame), involving one day or one week of GNSS observations. It can of course be applied to any kind of network (being global or regional) not only for positions, but also for velocities. It comprises the following steps:

1. Selection of a reference set of known ITRF/IGS stations and collecting their GNSS observation data provided in Receiver INdependent EXchange (RINEX) format from IGS data centers, covering the timespan (one day or one week) of the implied observations. It is highly advised to select a set of ITRF/IGS stations that are as homogeneously and globally distributed as possible, in order to achieve the best and an accurate expression in the ITRF.
2. Processing user station data together with the selected ITRF/IGS ones, using the preferred GNSS software. In this step, IGS orbits, clocks and EOPs should be fixed to the values consistent with the associated ITRF/IGS frame (ITRFyy, IGSyy). Fixing or tightly constraining ITRF/IGS reference station coordinates should by all means be avoided. Doing so would potentially introduce distortion in the solution due to possible outdated ITRF/IGS station coordinates after some events, such as earthquakes or equipment changes. Moreover, as the ITRF is

a secular linear frame, fixing or tightly constraining ITRF/IGS reference station coordinates would also inhibit the geophysical signal embedded in the transformed solution one may want to preserve.

3. Propagation of the selected ITRF/IGS station positions (X_I) at the central epoch (t_c) of the employed GNSS observations, using

$$X_I(t_c) = X_I(t_0) + \dot{X}_I \cdot (t_c - t_0),$$

where $X_I(t_0)$ are the ITRF/IGS station positions at epoch t_0 and \dot{X}_I are their linear velocities.

4. Application of minimum constraints approach [36.34, 35], detailed below, which is believed to be implemented in all major scientific software packages. The derived solution will be expressed in the ITRF/IGS frame that is consistent with the used orbits.
5. Comparison of the estimated ITRF/IGS reference station positions to the official published values, propagated at epoch t_c in step 3, by fitting a similarity transformation of three, four or seven parameters selected in the MC application and checking for consistency. The estimated transformation parameters should all be zero. In addition, if large discrepancies (postfit residuals of the similarity transformation) are found for some stations (exceeding a certain threshold, say 1–2 cm, but depending on the targeted accuracy), these stations should be rejected from the ITRF/IGS reference set and the processing chain should be iterated. Care should also be taken in the time interval of the validity of the used IGS/ITRF coordinates, taking into account station position discontinuities.

The starting point of the MC concept is based on the seven-parameter similarity transformation between any two reference systems or frames. Therefore, the linearized relationship between any space geodesy TRF solution, for example GNSS-based solution (X_G) and the ITRF (X_I), over selected reference set of common stations, can be written as

$$X_I = X_G + A\theta, \quad (36.7)$$

where the design matrix A is a stacked matrix made from elementary 3-row matrices

$$A^i = \begin{pmatrix} 1 & 0 & 0 & x_a^i & 0 & z_a^i & -y_a^i \\ 0 & 1 & 0 & y_a^i & -z_a^i & 0 & x_a^i \\ 0 & 0 & 1 & z_a^i & y_a^i & -x_a^i & 0 \end{pmatrix} \quad (36.8)$$

with $i = 1 \dots n$ for a total of n sites, and where $\theta = (T_x, T_y, T_z, D, R_x, R_y, R_z)^T$ is the vector of seven transformation parameters. T_x, T_y, T_z are the three translation

components, D is the scale factor, and R_x, R_y, R_z are the three rotation parameters. The approximate coordinates x_a^i, y_a^i, z_a^i of point i , appearing in the design matrix \mathbf{A} can be taken from the ITRF/IGS reference solution. Note that (36.7) is only valid at the same and common epoch of the two station position sets (\mathbf{X}_I and \mathbf{X}_G). It can also be generalized to 14 parameters when station velocities are involved in the process; see [36.24] for more details. Note also that the design matrix \mathbf{A} can be reduced to the columns corresponding to the frame parameters of interest, e.g., columns 1, 2 and 3 for the origin components; 5, 6 and 7 for the orientation parameters. In case of a regional network, applying the MC approach on the three translation components (i.e., \mathbf{A} is reduced to the three first columns) can be sufficient. It is however advisable to evaluate at least the following three options: translation, translation and scale, all seven parameters.

The unweighted least squares expression of (36.7) yields for θ

$$\theta = \overbrace{(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top}^{\mathbf{B}} (\mathbf{X}_I - \mathbf{X}_G). \quad (36.9)$$

The approach of MC consists in using the matrix $\mathbf{B} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top$ in such a way that \mathbf{X}_G will be expressed in the same frame as the ITRF solution \mathbf{X}_I . Therefore to have \mathbf{X}_G expressed in the ITRF at a certain Σ_θ level, an MC equation can be written as

$$\mathbf{B}(\mathbf{X}_I - \mathbf{X}_G) = \mathbf{0}(\Sigma_\theta), \quad (36.10)$$

where Σ_θ is the variance matrix at which (36.10) is satisfied. It is a diagonal matrix containing small variances (to be selected at the user level) for each one of the seven transformation parameters. It is suggested to use 0.1 mm for translation parameters and equivalent amounts (i.e., 0.1 mm divided by the Earth radius) for the scale and orientation parameters.

In terms of normal equations, we can then write

$$\mathbf{B}^\top \Sigma_\theta^{-1} \mathbf{B}(\mathbf{X}_I - \mathbf{X}_G) = \mathbf{0}. \quad (36.11)$$

The initial normal equation system of a GNSS-based solution before adding any kind of constraints can be written as

$$\mathbf{N}(\Delta \mathbf{X}) = \mathbf{K}, \quad (36.12)$$

where $\Delta \mathbf{X} = \mathbf{X} - \mathbf{X}_{\text{apr}}$, with \mathbf{X} being the unknown vector, \mathbf{X}_{apr} is the vector of a priori values, \mathbf{N} is the unconstrained normal matrix and \mathbf{K} is the right-hand side vector.

By fixing the IGS products (orbits, clocks and EOPs), the normal equation system (36.12) becomes invertible, but the underlying TRF could be far from that

of the ITRF, i.e., defined at the level of the orbit precision (a few cm). The same normal equation system can also be obtained after removing classical constraints applied to a given GNSS-based solution.

Selecting a subset of ITRF stations (\mathbf{X}_I), the MC equation becomes

$$\mathbf{B}^\top \Sigma_\theta^{-1} \mathbf{B}(\Delta \mathbf{X}) = \mathbf{B}^\top \Sigma_\theta^{-1} \mathbf{B}(\mathbf{X}_I - \mathbf{X}_{\text{apr}}). \quad (36.13)$$

Note that the right-hand side of (36.13) vanishes if the a priori values are taken from the ITRF/IGS selected solution.

Cumulating (36.12) and (36.13) yields

$$\begin{aligned} (\mathbf{N} + \mathbf{B}^\top \Sigma_\theta^{-1} \mathbf{B})(\Delta \mathbf{X}) \\ = \mathbf{K} + \mathbf{B}^\top \Sigma_\theta^{-1} \mathbf{B}(\mathbf{X}_I - \mathbf{X}_{\text{apr}}). \end{aligned} \quad (36.14)$$

The minimally constrained solution, expressed in the ITRF upon the selected stations is then

$$\begin{aligned} \mathbf{X} &= \mathbf{X}_{\text{apr}} + (\mathbf{N} + \mathbf{B}^\top \Sigma_\theta^{-1} \mathbf{B})^{-1} \\ &\quad \times (\mathbf{K} + \mathbf{B}^\top \Sigma_\theta^{-1} \mathbf{B}(\mathbf{X}_I - \mathbf{X}_{\text{apr}})). \end{aligned} \quad (36.15)$$

36.2.5 GNSS, Reference Frame and Sea Level Monitoring

Because of its ramifications around climate change and global warming, sea level monitoring requires the most stringent continuous geodetic observations that can only be addressed within the context of a global and stable reference frame. Two main data streams are used to infer sea level rise and its spatial and temporal variability: tide gage records and satellite altimetry data. The former dataset requires precise quantification of land vertical motion where tide gages are located, and the latter dataset imposes the precise knowledge of satellite orbits in a well-defined global reference frame. Both methods greatly benefit from the availability of GNSS observations. GNSS is the technique of choice to infer vertical crustal motion due to its ease of use and its connection to the ITRF through IGS products. Data collected by GPS receivers on board altimetry satellites and by ground-based receivers, together with DORIS and SLR data are used to precisely determine satellite orbits [36.59, 60].

To fully exploit tide gage records and accurately determine land vertical motion, it has been demonstrated that the GNSS processing strategy, together with the availability of an accurate reference frame are the main two limiting factors for improving our understanding of regional sea level variability in space and time. The processing strategy includes precise orbit determination of the GNSS satellites, an optimal

treatment of GNSS terrestrial observations and an advanced method of reference frame determination. Using an improved processing strategy and GPS data spanning 10 y at tide gages, [36.61] determined vertical velocities, based on ITRF2005, with uncertainties several times smaller than the 1–3 mm/y associated with global sea level change. The same authors have also shown that GPS-based land motion corrections at tide gages and expressed in ITRF2005 perform much better than glacial isostatic adjustment (GIA) model predictions, both on the global and the regional scale. These results suggest that GNSS measurements are more appropriate than GIA models to capture localized vertical motions associated with, for example, plate tectonics, volcanism, sediment compaction, or underground fluid extraction.

Precise orbit determination (POD) is one of the main critical issues for an accurate determination of global sea level variability using altimetry data of ocean surface satellite topography missions (OSTM), such as TOPEX-Poseidon, and Jason-1 and 2. These missions carry onboard three tracking systems (DORIS,

GPS and SLR) to meet the requirement of better than 1.5 cm radial accuracy for the operational orbit included in the geophysical data record products [36.60]. One of the main long-period error sources of POD is the stability of the origin of the reference frame, and in particular its z -component. In [36.62] it was shown that the difference between using an old reference frame called CSR95 (compatible with ITRF2000) of the Center for Space Research of the University of Texas at Austin, and ITRF2005 in orbit computation caused a change in the estimated mean sea level trend of -0.26 mm/y for the period from 1993 to 2002. The primary cause was shown to be the drift in the z -component of the origin between the two frames of 1.8 mm/y that also affected the regional sea level rates at the high latitudes by ± 1.5 mm/y. The review article [36.60] gives a summary of the different levels of performance of POD estimates as determined by different groups, using data of the three satellite techniques. They reported in particular that the 1 cm goal is met by both Jason-1 and 2 GPS-based reduced-dynamic orbits.

36.3 Earth Rotation, Polar Motion, and Nutation

Observations of the Earth's rotation show that while the Earth rotates about its axis once a day, it does not do so uniformly. Instead, the rate of rotation fluctuates by as much as a millisecond a day. The Earth wobbles as it rotates because its mass is not balanced about its rotation axis, and the Earth precesses and nutates in space. These variations in the Earth's rotation are caused by processes acting within the interior of the Earth such as glacial isostatic adjustment and core-mantle interaction torques, by processes acting at the surface of the Earth such as fluctuations in the transport of mass within the atmosphere and oceans, and by processes acting external to the Earth such as torques due to the gravitational attraction of the Sun, Moon, and planets [36.63–67].

In principle, only three time-dependent parameters, the Euler angles, are needed to fully characterize the varying orientation of the Earth in space. However, by convention, five parameters are actually used: two precession and nutation parameters that give the location of the reference pole in the space-fixed celestial frame, two polar motion parameters that give the location of the reference pole in the body-fixed terrestrial frame, and a spin parameter that gives the angular rotation of the Earth about the reference axis. The advantage of using five parameters instead of three is that with five parameters the externally forced precession/nutation motion of

the Earth is largely separated from its internally excited wobbling motion, also known as polar motion.

Routine measurements of the Earth's time-varying rotation are currently provided by the space-geodetic techniques of satellite and lunar laser ranging (SLR and LLR), very long baseline interferometry (VLBI), global navigation satellite systems (GNSSs) like the global positioning system (GPS), and Doppler orbitography and radio positioning integrated by satellite (DORIS). Each of these techniques has its own unique strengths and weaknesses in its ability to determine the five Earth orientation parameters (EOPs). Not only is each technique sensitive to a different subset and/or linear combination of the Earth orientation parameters, but also the averaging time for their determination is different, as is the interval between observations and the precision with which they can be determined.

Because of the large number of Earth orbiting satellites that transmit GNSS signals and the large number of ground stations that receive them, continuous, uninterrupted measurements of the Earth's rotation are provided by GNSS. In addition, because the raw observables can be rapidly analyzed, GNSS can provide measurements of the Earth's rotation in near-real time. In this section, the contribution of GNSS to monitoring the rotational behavior of the Earth and to understanding the causes of the observed variations is discussed.

36.3.1 Theory of the Earth's Rotation

Changes in the rotation of the solid Earth are usually studied by applying the principle of conservation of angular momentum, which requires that changes in the rotation vector of the solid Earth are manifestations of either torques acting on the solid Earth or of changes in the mass distribution within the solid Earth that alter its inertia tensor. Angular momentum is transferred between the solid Earth and the fluid regions (the underlying liquid metallic core and the overlying hydrosphere and atmosphere) with which it is in contact; concomitant torques are due to hydrodynamic or magnetohydrodynamic stresses acting at the fluid/solid Earth interfaces. Using the principle of the conservation of angular momentum the equations governing small variations in both the rate of rotation and in the position of the rotation vector with respect to the Earth's crust can be derived [36.67–70].

Within a rotating, body-fixed reference frame, the equation that relates changes in the angular momentum $\mathbf{L}(t)$ of a rotating body to the external torques $\boldsymbol{\tau}(t)$ acting on the body is [36.71]

$$\frac{\partial}{\partial t} [\mathbf{h}(t) + \mathbb{I}(t) \cdot \boldsymbol{\omega}(t)] + \boldsymbol{\omega}(t) \times [\mathbf{h}(t) + \mathbb{I}(t) \cdot \boldsymbol{\omega}(t)] = \boldsymbol{\tau}(t), \quad (36.16)$$

where $\boldsymbol{\omega}(t)$ is the angular velocity of the body with respect to inertial space and where $\mathbf{L}(t)$ has been written as the sum of two terms: (1) that part $\mathbf{h}(t)$ due to motion relative to the rotating reference frame, and (2) that part due to changes in the inertia tensor $\mathbb{I}(t)$ of the body caused by changes in the distribution of mass.

The Earth's rotation deviates only slightly from a state of uniform rotation, the deviation being a few parts in 10^8 in speed, corresponding to changes of a few milliseconds (ms) in the length of the day, and about a part in 10^6 in the orientation of the rotation axis relative to the crust of the Earth, corresponding to a variation of several hundred milliarcseconds (mas) in polar motion. Such small deviations in rotation are studied by linearizing (36.16).

Let the Earth be initially uniformly rotating about its figure axis and orient the body-fixed reference frame so that its z -axis is aligned with the figure axis. Under a small perturbation to this initial state, the initial relative angular momentum \mathbf{h}_0 (which is zero because there is initially no relative angular momentum) will be perturbed to $\mathbf{h}_0 + \Delta\mathbf{h}$, the initial inertia tensor \mathbb{I}_0 will be perturbed to $\mathbb{I}_0 + \Delta\mathbb{I}$, and the initial angular velocity vector $\boldsymbol{\omega}_0$ will be perturbed to $\boldsymbol{\omega}_0 + \Delta\boldsymbol{\omega}$. Keeping terms to first order in small quantities and making a number of other assumptions including assuming that the

Earth is axisymmetric, that the oceans remain in equilibrium as the rotation of the solid Earth changes, that the core is not coupled to the mantle, and that the rotational variations occur on timescales much longer than a day, then the Cartesian components of the linearized version of (36.16) can be written as [36.67]

$$\frac{1}{\sigma_o} \frac{\partial m_x(t)}{\partial t} + m_y(t) = \chi_y(t) - \frac{1}{\Omega} \frac{\partial \chi_x(t)}{\partial t} \quad (36.17)$$

$$\frac{1}{\sigma_o} \frac{\partial m_y(t)}{\partial t} - m_x(t) = -\chi_x(t) - \frac{1}{\Omega} \frac{\partial \chi_y(t)}{\partial t} \quad (36.18)$$

$$m_z(t) = -\chi_z(t), \quad (36.19)$$

where Ω is the mean angular velocity of the Earth (rad/s), σ_o is the observed complex-valued frequency of the Chandler wobble (rad/s), and the dimensionless m_i are related to the elements of the perturbed rotation vector

$$\begin{aligned} \boldsymbol{\omega}(t) &= \boldsymbol{\omega}_o(t) + \Delta\boldsymbol{\omega}(t) \\ &= \Omega \hat{\mathbf{z}} + \Omega [m_x(t) \hat{\mathbf{x}} + m_y(t) \hat{\mathbf{y}} + m_z(t) \hat{\mathbf{z}}], \end{aligned} \quad (36.20)$$

with the hat denoting a vector of unit length.

The dimensionless $\chi_i(t)$ in (36.17)–(36.19) are known as excitation functions and are functions of the perturbed inertia tensor ($\text{kg}\cdot\text{m}^2$) and relative angular momentum ($\text{kg}\cdot\text{m}^2/\text{s}$) that are exciting the changes in the Earth's rotation [36.67]

$$\chi_x(t) = \frac{1.608[\Delta h_x(t) + 0.684 \Omega \Delta I_{xz}(t)]}{(C - A')\Omega}, \quad (36.21)$$

$$\chi_y(t) = \frac{1.608[\Delta h_y(t) + 0.684 \Omega \Delta I_{yz}(t)]}{(C - A')\Omega}, \quad (36.22)$$

$$\chi_z(t) = \frac{0.997}{C_m \Omega} [\Delta h_z(t) + 0.750 \Omega \Delta I_{zz}(t)], \quad (36.23)$$

where C is the axial principal moment of inertia of the entire Earth, C_m is that of just the crust and mantle, and A' is the average $(A+B)/2$ of the equatorial principal moments of inertia of the entire Earth.

The numerical coefficients of the inertia tensor terms in (36.21)–(36.23) are functions of load Love numbers [36.67], so (36.21)–(36.23) are valid for processes like atmospheric surface pressure variations that load the solid Earth causing it to deform. For processes that do not load the solid Earth, like earthquakes, the coefficients 0.684 in (36.21)–(36.22) and 0.750 in (36.23) should be set to 1.0.

36.3.2 Length-of-Day

Equations (36.19) and (36.23) relate changes in the axial component of the Earth's angular velocity to changes

in both the axial component of relative angular momentum and in the zz -element of the inertia tensor. But GNSS observations do not give changes in the axial component of the Earth's angular velocity. Instead, they give changes in the length of the day. The length of the day is the rotational period of the Earth. Changes $\Delta\Lambda(t)$ in the length of the day are related to the time rate-of-change of the difference (UT1 – TAI) between Universal Time UT1 and atomic time TAI and to changes $\Delta\omega_z(t) = \Omega m_z(t)$ in the axial component of the Earth's angular velocity [36.67]

$$\begin{aligned}\frac{\Delta\Lambda(t)}{\Lambda_0} &= -\frac{d(\text{UT1} - \text{TAI})}{dt} \\ &= -\frac{\Delta\omega_z(t)}{\Omega} = -m_z(t),\end{aligned}\quad (36.24)$$

where Λ_0 is the nominal length-of-day (LOD) of 86 400 s. GNSS-observed changes in the length of the day are therefore related to the processes causing the length-of-day to change by

$$\frac{\Delta\Lambda(t)}{\Lambda_0} = \frac{0.997}{C_m\Omega} [\Delta h_z(t) + 0.750 \Omega \Delta I_{zz}(t)].\quad (36.25)$$

Figure 36.5 shows the changes in the length of the day $\Delta\Lambda(t)$ measured by GNSS during March 1997 to June 2014. Like a spinning ice skater whose speed of rotation increases as the skater's arms are brought closer to the body, the speed of the Earth's rotation increases and the length of the day decreases if its mass is brought closer to its axis of rotation. Conversely, the speed of the

Earth's rotation decreases and the length of the day increases if its mass is moved away from the rotation axis. Observations of the length of the day like those shown in Fig. 36.5 show that it consists mainly of:

1. A linear trend of rate +1.8 ms/cy (not evident in Fig. 36.5 because of the shortness of the record)
2. Decadal variations having an amplitude of a few milliseconds
3. Tidal variations having an amplitude of about 1 ms
4. Seasonal variations having an amplitude of about 0.5 ms
5. Smaller amplitude variations occurring on all measurable timescales.

A number of different dynamical Earth processes are responsible for the changes in the length of the day shown in Fig. 36.5. Tidal forces due to the changing gravitational attraction of the Sun, Moon, and planets deform the solid and fluid regions of the Earth, causing the Earth's rotation to change by causing its inertia tensor to change. In fact, solid-body tides, caused by the tidal forces acting on the solid Earth, are the dominant cause of length-of-day variations on intraseasonal to interannual timescales. Ocean tides, caused by the tidal forces acting on the oceans, are the dominant cause of subdaily length-of-day variations and contribute to length-of-day variations at longer periods.

Figure 36.6 shows a spectrum of the GNSS-observed length-of-day variations that are shown in Fig. 36.5. Peaks at the tidal frequencies are clearly evident. Nontidal variations in the length of the day occurring on timescales of a few days to a few years are predominantly caused by variations in the zonal atmospheric winds, with variations in atmospheric surface pressure, oceanic currents and bottom pressure, and water stored on land contributing much less. On longer timescales, decadal variations as large as a few milliseconds in the length of the day are caused by interactions between the fluid outer core and solid mantle of the Earth, and a secular trend of +1.8 ms/cy in the length of the day is caused by a combination of tidal dissipation in the Earth-Moon system (+2.3 ms/cy) and by glacial isostatic adjustment (–0.5 ms/cy). See [36.67] for a review of these and other causes of length-of-day changes.

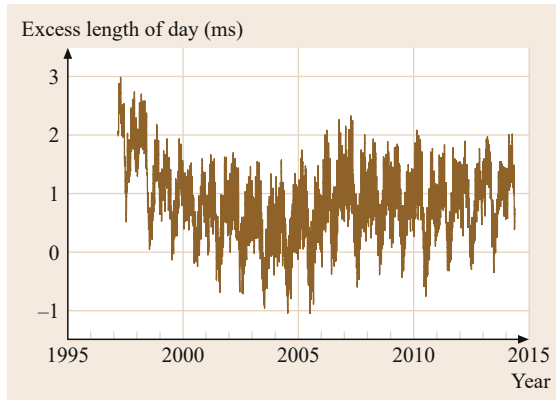


Fig. 36.5 Observed excess length-of-day values in milliseconds (ms) spanning March 1997 to June 2014 from the IGS final combined series. The excess length-of-day is the amount by which the length-of-day is longer (positive values) or shorter (negative values) than the nominal length-of-day of 86 400 s

36.3.3 Polar Motion

Equations (36.17)–(36.18) and (36.21)–(36.22) relate changes in the equatorial components of the Earth's angular velocity to changes in both the equatorial components of relative angular momentum and in the xz - and yz -elements of the inertia tensor. But GNSS observations do not give changes in the equatorial com-

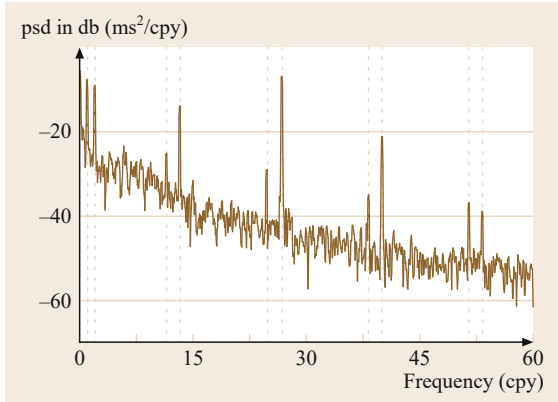


Fig. 36.6 Power spectral density (psd) estimates in decibels (db) computed by the multitaper method of the IGS Final combined length-of-day measurements spanning March 1997 to June 2014. Vertical dashed lines indicate the frequencies of the annual (1 cycle/y (cpy)) and semiannual (2 cpy) LOD variations and of the largest tidal variations in the monthly (13 cpy), fortnightly (27 cpy), termensual (40 cpy), and 7 d (51 cpy) tidal bands

ponents of the Earth's angular velocity. Instead, they give changes in the terrestrial position of the celestial intermediate pole (CIP). The CIP is the intermediate reference pole whose use allows the separation of nutation from wobble. The equatorial components of the Earth's angular velocity are related to the GNSS-observed position of the CIP in the terrestrial reference frame $\mathbf{p}(t) = p_x(t) - j p_y(t)$, where the negative sign accounts for $p_y(t)$ being conventionally positive toward 90° West longitude, by [36.67]

$$\mathbf{m}(t) = \mathbf{p}(t) - \frac{j}{\Omega} \frac{d\mathbf{p}(t)}{dt}, \quad (36.26)$$

where $\mathbf{m}(t) = m_x(t) + j m_y(t)$ and j is the imaginary unit $\sqrt{-1}$. By combining (36.26) with (36.17)–(36.18) it can be shown that the GNSS-observed polar motion parameters $p_x(t)$ and $p_y(t)$ are related to the polar motion excitation functions $\chi(t) = \chi_x(t) + j \chi_y(t)$ by

$$\begin{aligned} \mathbf{p}(t) + \frac{j}{\sigma_0} \frac{d\mathbf{p}(t)}{dt} &= \chi(t) \\ &= \frac{1.608[\Delta\mathbf{h}(t) + 0.684\Omega\Delta\mathbb{I}(t)]}{(C - A')\Omega}, \end{aligned} \quad (36.27)$$

where in this equation $\Delta\mathbf{h}(t) = \Delta h_x(t) + j \Delta h_y(t)$ and $\Delta\mathbb{I}(t) = \Delta I_{xz}(t) + j \Delta I_{yz}(t)$.

Figures 36.7a and 36.7b show the x - and y -components of polar motion, $p_x(t)$ and $p_y(t)$ respectively, measured by GNSS during July 1996 to June 2014. Much like the wobble of an unbalanced

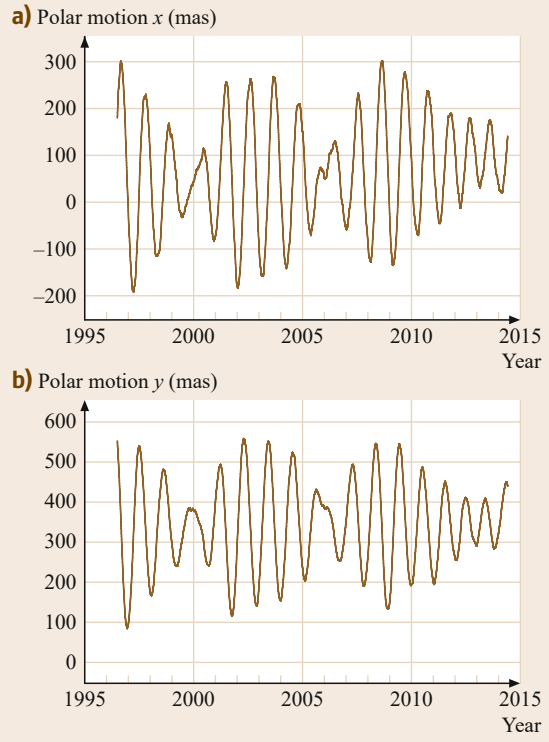


Fig. 36.7 The x -component (a) and y -component (b) of observed polar motion values in milliarcseconds (mas) spanning July 1996 to June 2014 from the IGS final combined series. The readily apparent beat pattern is caused by the 12-month annual and 14-month Chandler wobbles, which have similar amplitudes, constructively and destructively interfering with each other

automobile tire, the Earth wobbles because the mass of the Earth is not balanced about its rotation axis. In the absence of excitation, the Earth would eventually stop wobbling because of dissipation processes in the oceans and solid, but not rigid, crust and mantle. But as long as mass continues to be horizontally transported towards or away from the poles the Earth will continue to wobble. Observations like those shown in Figs. 36.7a and 36.7b show that polar motion consists mainly of:

1. A forced annual wobble having a nearly constant amplitude of about 100 mas
2. The free Chandler wobble having a period of about 433 d and a variable amplitude ranging from about 100–200 mas
3. Quasiperiodic variations on decadal timescales having amplitudes of about 30 mas known as the Markowitz wobble
4. A linear trend having a rate of about 3.5 mas/y and a direction towards 79° West longitude

5. Smaller amplitude variations occurring on all measurable timescales.

One of the great strengths of GNSS in measuring changes in the Earth's rotation is that it can measure not only the polar motion parameters themselves but also their time rate-of-change. By combining polar motion and polar motion-rate measurements according to the left-hand side of (36.27), GNSS allows direct measurements of the polar motion excitation functions to be made. These directly measured excitation functions can then be compared to models of, say, atmospheric and oceanic excitation to study the causes of the observed polar motion. Figure 36.8 shows a spectrum of the polar motion excitation functions determined from the GNSS-observed polar motion and polar motion-rate measurements. Like the changes in the length of the day, a number of different dynamical Earth processes are responsible for exciting polar motion. Because the tidal potential is symmetric about the polar axis, tidal deformations of the solid Earth do not cause it to wobble. But because ocean basins are asymmetrically distributed about the Earth, ocean tides do cause the Earth to wobble and small peaks at the fortnightly tidal frequencies are clearly evident in Fig. 36.8.

The annual wobble is a forced motion of the Earth that has been shown to be largely caused by the an-

nual appearance of a high atmospheric pressure system over Siberia every winter [36.63]. This Siberian high-pressure system annually loads the Siberian crust, causing the Earth to wobble with an annual period. The Chandler wobble on the other hand is not a forced motion of the Earth, but is instead a free resonant motion of the Earth that occurs because the Earth is not rotating about its figure axis, the axis about which the Earth's mass is balanced. The Chandler wobble would freely decay with an exponential time constant of about 68 y if no mechanism or mechanisms were acting to excite it. Using atmospheric and oceanic general circulation models, it has been shown that the sum of atmospheric surface pressure and ocean-bottom pressure variations are the primary source of excitation of the Chandler wobble, with ocean-bottom pressure variations being about twice as effective as atmospheric pressure variations over land. On the longest timescales, the trend in the pole path has been shown to be caused by a combination of the viscoelastic response of the Earth to past changes in ice sheet mass and the elastic response of the Earth to present-day changes [36.72]; see [36.67] for a review of these and other causes of polar motion.

36.3.4 Nutation

Because of their great distance, the radio reference sources observed by VLBI exhibit negligible motion in the sky and can therefore be used to realize an inertial, celestial reference frame. This allows VLBI to determine all five of the Earth orientation parameters that are conventionally used to fully characterize the orientation of the Earth in space, including the two nutation parameters. But because the large nongravitational forces acting on artificial Earth-orbiting satellites cannot be accurately modeled [36.54], the orbits of satellites cannot be used to realize an inertial reference frame. Thus satellite techniques like GNSS can determine only a subset of the five EOPs. In particular, because of correlations between satellite orbital elements and both UT1 and the two nutation parameters, these Earth orientation parameters cannot be determined by satellite techniques like GNSS. However, their rate-of-change can be determined.

The rate-of-change of UT1, or length-of-day (36.24), was first routinely estimated from GPS data in June 1992 by the Center for Orbit Determination in Europe (CODE) analysis center. In [36.73] it was subsequently argued that there was no fundamental difference between estimating rates in UT1 and rates in nutation and that consequently GNSS should also be able to measure the rate-of-change of nutation.

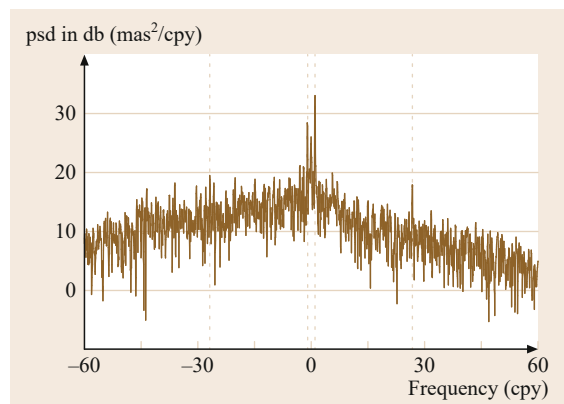


Fig. 36.8 Power spectral density (psd) estimates in decibels (db) computed by the multitaper method of the polar motion excitation functions $\chi(t)$ spanning July 1996 to June 2014 formed by using (36.27) to combine the IGS final combined polar motion and polar-motion-rate measurements. Vertical dashed lines indicate the prograde and retrograde frequencies of the annual excitation (± 1 cycle/y (cpy)) and of the fortnightly tidal term (± 27 cpy). The retrograde component of polar motion excitation is represented by negative frequencies, the prograde component by positive frequencies

They showed that the uncertainty in GNSS-measured nutation rates should grow linearly with the period of the nutation term and that GNSS should therefore be able to measure the rates of nutation terms having short periods. Using 3.5 y of GNSS data they were able to estimate the rates of 34 nutation terms having periods between four and 16 d.

Acknowledgments. The work of Zuheir Altamimi described in this chapter was performed at IGN France, host of the ITRF Center. The work of Richard Gross described in this chapter was performed at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

- 36.1 B. Hofmann-Wellenhof, H. Moritz: *Physical Geodesy*, 2nd edn. (Springer, Vienna 2006)
- 36.2 W. Torge, J. Müller: *Geodesy*, 4th edn. (De Gruyter, Berlin 2012)
- 36.3 R. Rummel: Global integrated geodetic and geodynamic observing system (GIGGOS). In: *Towards an Integrated Global Geodetic Observing System (IG-GOS)*, IAG Symposia, Vol. 120, ed. by R. Rummel, H. Drewes, W. Bosch, H. Hornik (Springer, Berlin 2000) pp. 253–260
- 36.4 H.-P. Plag, G. Beutler, R. Gross, T.A. Herring, C. Rizos, R. Rummel, D. Sahagian, J. Zumberge: Introduction. In: *Global Geodetic Observing System: Meeting the Requirements of a Global Society on a Changing Planet in 2020*, ed. by H.-P. Plag, M. Pearlman (Springer, Berlin 2009) pp. 1–13
- 36.5 H. Drewes, H. Hornik, J. Ádám, S. Rózsa: The geodesist's hand book 2012, *J. Geod.* **86**(10), 787–974 (2012)
- 36.6 H.-P. Plag, M. Pearlman (Eds.): *Global Geodetic Observing System: Meeting the Requirements of a Global Society on a Changing Planet in 2020* (Springer, Berlin 2009)
- 36.7 G.W. Withee, D.B. Smith, M.B. Hales: Progress in multilateral earth observation cooperation: CEOS, IGOS, and the ad hoc group on earth observations, *Space Policy* **20**, 37–43 (2004)
- 36.8 G. Petit, B. Luzum: (*Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt 2010*), *IERS Technical Note*, IERS Conventions, Vol. 36, 2010
- 36.9 Z. Altamimi, L. Métivier, X. Collilieux: ITRF2008 plate motion model, *J. Geophys. Res.* **117**(B07402), 1–14 (2012)
- 36.10 J. Ray (Ed.): *IERS Technical, IERS Analysis Campaign to Investigate Motions of the Geocenter*, Vol. 25 (Central Bureau of IERS, Observatoire de Paris, Paris 1999) p. 121
- 36.11 X. Wu, J. Ray, T. van Dam: Geocenter motion and its geodetic and geophysical implications, *J. Geodyn.* **58**, 44–61 (2012)
- 36.12 L. Métivier, M. Greff-Lefftz, Z. Altamimi: On secular geocenter motion: The impact of climate changes, *Earth Planet. Sci. Lett.* **296**(3/4), 360–366 (2010)
- 36.13 E. Pavlis: Fortnightly resolution geocenter series: A combined analysis of Lageos 1, 2 SLR data (1993–1996). In: *IERS Analysis Campaign to Investigate Motions of the Geocenter*, ed. by J. Ray (Observatoire de Paris, Paris 1999) pp. 75–84
- 36.14 X. Collilieux, Z. Altamimi, J. Ray, T. van Dam, X. Wu: Effect of the satellite laser ranging network distribution on geocenter motion estimation, *J. Geophys. Res.* **114**(B04402), 1–17 (2009)
- 36.15 U. Hugentobler, H. van der Marel, T. Springer: Identification, mitigation of GNSS errors, *Proc. IGS Workshop 2006 Darmstadt*, ed. by T. Springer, G. Gendt, J.M. Dow (IGS, Pasadena 2006)
- 36.16 M. Meindl, G. Beutler, D. Thaller, R.R. Dach, A. Jäggi: Geocenter coordinates estimated from GNSS data as viewed by perturbation theory, *Adv. Space Res.* **51**(7), 1047–1064 (2013)
- 36.17 P. Rebischung, Z. Altamimi, T. Springer: A collinearity diagnosis of the GNSS geocenter determination, *J. Geod.* **88**, 65–85 (2014)
- 36.18 G. Blewitt, D. Lavallée, P. Clarke, K. Nurutdinov: A new global mode of Earth deformation: Seasonal cycle detected, *Science* **294**(5550), 2342–2345 (2001)
- 36.19 X. Wu, D.F. Argus, M.B. Heflin, E.R. Ivins, F.H. Webb: Site distribution and aliasing effects in the inversion for load coefficients and geocenter motion from GPS data, *Geophys. Res. Lett.* **29**(24), 63–1–63–4 (2002)
- 36.20 J. Ray, Z. Altamimi, X. Collilieux, T. van Dam: Anomalous harmonics in the spectra of GPS position estimates, *GPS Solutions* **12**, 55–64 (2008)
- 36.21 J. Griffiths, J. Ray: Sub-daily alias and draconitic errors in the IGS orbits, *GPS Solutions* **17**, 413–422 (2012)
- 36.22 Z. Altamimi: ITRF and co-location sites, *Proc. IERS Workshop Site Co-Location*, ed. by B. Richter, W.R. Dick, W. Schwegmann (2005), IERS Technical Note No. 33
- 36.23 R. Schmid, P. Steigenberger, G. Gendt, M. Ge, M. Rothacher: Generation of a consistent absolute phase-center correction model for GPS receiver and satellite antennas, *J. Geod.* **81**, 781–798 (2007)
- 36.24 Z. Altamimi, P. Sillard, C. Boucher: ITRF2000: A new release of the international terrestrial reference frame for Earth science applications, *J. Geophys. Res.* **107**(B10), 2214 (2002)
- 36.25 Z. Altamimi, X. Collilieux, J. Legrand, B. Garayt, C. Boucher: ITRF2005: A new release of the international terrestrial reference frame based on time series of station positions and Earth orientation parameters, *J. Geophys. Res.* **112**(B09401), 1–19 (2007)
- 36.26 Z. Altamimi, X. Collilieux, L. Métivier: ITRF2008: An improved solution of the international terrestrial

- reference frame, *J. Geod.* **85**(8), 457–473 (2011)
- 36.27 G. Blewitt, Y. Bock, J. Kouba: Constraining the IGS polyhedron by distributed processing, *Proc. IGS Workshop Densif. ITRF Reg. GPS Netw.*, Pasadena, ed. by J.F. Zumberge, R. Liu (1994) pp. 21–37
- 36.28 Z. Altamimi, D. Gambis, C. Bizouard: Rigorous combination to ensure ITRF and EOP consistency, *Proc. Journées 2007 Celest. Ref. Frame Futur.*, Meudon (Observatoire de Paris, Paris 2008) pp. 151–154
- 36.29 R. Neilan, J.M. Dow, G. Gendt: The international GPS service (IGS): Celebrating the 10th anniversary and looking to the next decade, *Adv. Space Res.* **36**(3), 320–326 (2005)
- 36.30 M.R. Pearlman, J.J. Degnan, J.M. Bosworth: The international laser ranging service, *Adv. Space Res.* **30**(2), 135–143 (2002)
- 36.31 P. Willis, H. Fagard, P. Ferrage, F.G. Lemoine, C.E. Noll, R. Noomen, M. Otten, J.C. Ries, M. Rothacher, L. Soudarin, G. Tavernier, J.J. Valette: The international DORIS service, toward maturity, *Adv. Space Res.* **45**(12), 1408–1420 (2010)
- 36.32 W. Schlüter, E. Himwich, A. Nothnagel, N. Vandenberg, A. Whitney: IVS and its important role in the maintenance of the global reference systems, *Adv. Space Res.* **30**(2), 145–150 (2002)
- 36.33 R. Ferland, M. Piraszewski: The IGS-combined station coordinates, Earth rotation parameters and apparent geocenter, *J. Geod.* **83**(3/4), 385–392 (2009)
- 36.34 Z. Altamimi, C. Boucher, P. Sillard: New trends for the realization of the international terrestrial reference system, *Adv. Space Res.* **30**(2), 175–184 (2002)
- 36.35 Z. Altamimi, A. Dermanis: The choice of reference system in ITRF formulation, *Proc. 7th Hotine-Marussi Symp. Mathem. Geod.*, Int. Assoc. Geod., Vol. 137, ed. by N. Sneeuw, P. Novák, M. Crespi, F. Sansò (Springer, Berlin, Heidelberg 2012) pp. 329–334
- 36.36 Z. Altamimi, C. Boucher, L. Duhem: The worldwide centimetric terrestrial reference frame and its associated velocity field, *Adv. Space Res.* **13**(11), 151–160 (1993)
- 36.37 J. Kouba, Y. Mireault, G. Beutler, T. Springer: A discussion of IGS solutions and their impact on geodetic and geophysical applications, *GPS Solutions* **2**(2), 3–15 (1998)
- 36.38 J. Kouba, Y. Mireault: 1998 Analysis Coordinator Report. In: *1998 Technical Reports*, ed. by K. GOWEY, R.E. Neilan, A. Moore (IGS Central Bureau, Jet Propulsion Laboratory, Pasadena 1999) pp. 15–58
- 36.39 J. Kouba, P. Héroux: Precise point positioning using IGS orbit and clock products, *GPS Solutions* **5**(2), 12–28 (2001)
- 36.40 J. Kouba: The GPS toolbox ITRF transformations, *GPS Solutions* **5**(3), 88–90 (2002)
- 36.41 R. Ferland: Reference frame working group technical report. In: *IGS 2001–2002 Technical Reports*, ed. by K. GOWEY, R. Neilan, A. Moore (JPL, Pasadena 2004) pp. 25–33
- 36.42 J. Ray: Reinforcing and securing the IGS reference tracking network, *Proc. Workshop State GPS Vert. Position. Precis.*: Sep. Earth Process. Space Geod., Cahiers du Centre Européen de Géodynamique et de Séismologie, Vol. 23, ed. by T. van Dam, O. Francis (Centre Européen de Géodynamique et de Séismologie, Luxembourg 2004) pp. 1–15
- 36.43 J. Ray, D. Dong, Z. Altamimi: IGS reference frames: Status and future improvements, *GPS Solutions* **8**(4), 251–266 (2004)
- 36.44 R. Schmid, M. Rothacher, D. Thaler, P. Steigenberger: Absolute phase center corrections of satellite and receiver antennas, *J. Geod.* **81**, 781–798 (2007)
- 36.45 P. Rebischung, J. Griffiths, J. Ray, R. Schmid, X. Collilieux, B. Garayt: IGS08: The IGS realization of ITRF2008, *GPS Solutions* **16**(4), 483–494 (2012)
- 36.46 R.F. Wong, C.M. Rollins, C.F. Minter: Recent updates to the WGS 84 reference frame, *Proc. ION GNSS*, Nashville (ION, Virginia 2012) pp. 1164–1172
- 36.47 V. Vdovin, A. Dorofeeva: Global geocentric coordinate system of the Russian federation, *Proc. 7th Meet. Int. Comm. GNSS (ICG)*, Work. Group D, Beijing (UNOOSA, Vienna 2012) pp. 1–15
- 36.48 Y. Yang: Chinese geodetic coordinate system 2000, *Chin. Sci. Bull.* **54**(15), 2714–2721 (2009)
- 36.49 G. Gendt, Z. Altamimi, R. Dach, W. Söhne, T. Springer: GGSP: Realisation, maintenance of the Galileo terrestrial reference frame, *Adv. Space Res.* **47**(2), 174–185 (2010)
- 36.50 Y. Hiyama, A. Yamagiwa, T. Kawahara, M. Iwata, Y. Fukuzaki, Y. Shouji, Y. Sato, T. Yutsudo, T. Sasaki, H. Shigematsu, H. Yamao, T. Inukai, M. Ohtaki, K. Kokado, S. Kurihara, I. Kimura, T. Tsutsumi, T. Yahagi, Y. Furuya, I. Kageyama, S. Kawamoto, K. Yamaguchi, H. Tsuji, S. Matsumura: Revision of survey results of control points after the 2011 off the Pacific coast of Tohoku earthquake, *Bull. Geospatial Inf. Auth. Jpn.* **59**, 31–42 (2011)
- 36.51 O. Montenbruck, P. Steigenberger, A. Hauschild: Broadcast versus precise ephemerides: A multi-GNSS perspective, *GPS Solutions* **19**(2), 321–333 (2015)
- 36.52 M.R. Craymer: The evolution of NAD83 in Canada, *Geomatica* **60**(2), 151–164 (2006)
- 36.53 T. Soler, R. Snay: Transforming positions and velocities between the international terrestrial reference frame of 2000 and North American datum of 1983, *J. Surv. Eng.* **130**(2), 130–249 (2004)
- 36.54 A. Milani, A.M. Nobili, P. Farinella: *Non-Gravitational Perturbations and Satellite Geodesy* (Adam Hilger, Bristol 1987)
- 36.55 J. Dawson, A. Woods: ITRF to GDA94 coordinate transformations, *J. Appl. Geod.* **4**, 189–199 (2010)
- 36.56 Z. Altamimi: ETRS89 realization: Current status, ETRF2005 and future development, *Bull. Geod. Geom.* **LXVIII**(3), 255–267 (2009)
- 36.57 G. Blick, C. Crook, D. Grant, J. Beavan: Implementation of a semi-dynamic datum for New Zealand. In: *A Window to the Future of Geodesy*, ed. by F. Sansò, Int. Assoc. Geod. Symp. Ser., Vol. 128 (2005) pp. 38–43
- 36.58 Z. Altamimi, P. Sillard, C. Boucher: ITRF2000: From theory to implementation, 5th Hotine-Marussi Symp. Mathem. Geod., Int. Assoc. Geod., Vol. 127, ed. by F. Sansò (2004) pp. 157–163

- 36.59 F.G. Lemoine, N.P. Zelensky, D.S. Chinn, D.E. Pavlis, D.D. Rowlands, B.D. Beckley, S.B. Luthcke, P. Willis, M. Ziebart, A. Sibthorpe, J.P. Boy, V. Luceri: Towards development of a consistent orbit series for TOPEX, Jason-1, and Jason-2, *Adv. Space Res.* **46**(12), 1513–1540 (2010)
- 36.60 L. Cerri, J.P. Berthias, W.I. Bertiger, B.J. Haines, F.G. Lemoine, F. Mercier, J.C. Ries, P. Willis, N.P. Zelensky, M. Ziebart: Precision orbit determination standards for the Jason series of altimeter missions, *Mar. Geod.* **33**(S1), 379–418 (2010)
- 36.61 G. Wöppelmann, C. Letetrel, A. Santamaria, M.-N. Bouin, X. Collilieux, Z. Altamimi, S.D.P. Williams, B. Martin Miguez: Rates of sea-level change over the past century in a geocentric reference frame, *Geophys. Res. Lett.* **36**(L12607), 1–6 (2009)
- 36.62 B.D. Beckley, F.G. Lemoine, S.B. Luthcke, R.D. Ray, N.P. Zelensky: A reassessment of TOPEX and Jason-1 altimetry based on revised reference frame and orbits, *Geophys. Res. Lett.* **34**(L14608), 1–5 (2007)
- 36.63 W.H. Munk, G.J.F. MacDonald: *The Rotation of the Earth: A Geophysical Discussion* (Cambridge Univ. Press, New York 1960)
- 36.64 K. Lambeck: *The Earth's Variable Rotation: Geophysical Causes and Consequences* (Cambridge Univ. Press, New York 1980)
- 36.65 K. Lambeck: *Geophysical Geodesy: The Slow Deformations of the Earth* (Oxford Univ. Press, New York 1988)
- 36.66 T.M. Eubanks: Contributions of space geodesy to geodynamics: Earth dynamics. In: *Variations in the Orientation of the Earth*, ed. by D.E. Smith, D.L. Turcotte (American Geophysical Union, Washington DC 1993) pp. 1–54
- 36.67 R.S. Gross: Earth rotation variations – Long period. In: *Physical Geodesy*, ed. by T.A. Herring (Elsevier, Oxford 2007) pp. 239–294
- 36.68 M.L. Smith, F.A. Dahlen: The period and Q of the Chandler wobble, *Geophys. J. Roy. Astron. Soc.* **64**, 223–281 (1981)
- 36.69 J.M. Wahr: The effects of the atmosphere, oceans on the Earth's wobble – I. Theory, *Geophys. J. Roy. Astron. Soc.* **70**, 349–372 (1982)
- 36.70 J.M. Wahr: The effects of the atmosphere and oceans on the Earth's wobble and on the seasonal variations in the length of day – II. Results, *Geophys. J. Roy. Astron. Soc.* **74**, 451–487 (1983)
- 36.71 H. Goldstein: *Classical Mechanics* (Addison-Wesley, Reading 1950)
- 36.72 J.L. Chen, C.R. Wilson, J.C. Ries, B.D. Tapley: Rapid ice melting drives Earth's pole to the east, *Geophys. Res. Lett.* **40**, 2625–2630 (2013)
- 36.73 M. Rothacher, G. Beutler, T.A. Herring, R. Weber: Estimation of nutation using the global positioning system, *J. Geophys. Res.* **104**(B3), 4835–4859 (1999)

Geodynamics

37. Geodynamics

Jeff Freymueller

Geodynamic studies rely on measurement of motions over time, such as displacements, displacement time series, or velocities for those sites that move steadily with time. Global navigation satellite systems (GNSSs) are widely used for geodynamics research, including studies of tectonic plate motions and plate boundary deformation, earthquakes and seismology, volcano deformation, surface loading deformation, and glacial isostatic adjustment. GNSS is an ideal tool for these studies because it can provide time series of millimeter-precision positions using inexpensive, portable and easily deployed equipment. This chapter illustrates and summarizes the important concepts and the basic computational models used to relate active processes within the Earth to surface deformation that can be observed using GNSS. These include conceptual models for the earthquake cycle, elastic dislocation theory, the Mogi volcanic source model, and surface loading computations. The chapter also summarizes important research results in all of these topics. Rapid and real-time applications of GNSS to use surface deformation for earthquake and tsunami warning are growing, and are likely to become even more important in the future, as will multi-GNSS observations to provide greater measurement accuracy.

37.1	GNSS for Geodynamics	1064
37.1.1	Accuracy Requirements	1064
37.1.2	Today's GNSS Accuracy	1065
37.1.3	Accuracy Limitations and Error Sources	1066
37.2	History and Establishment of GNSS Networks for Geodynamics	1067
37.2.1	Campaign GPS Networks	1067
37.2.2	Continuous GNSS Networks for Geodynamics	1068
37.2.3	The Importance of Global Networks	1071
37.3	Rigid Plate Motions	1071
37.4	Plate Boundary Deformation and the Earthquake Cycle	1073
37.4.1	Plate Boundary Zones	1074
37.4.2	Earthquake Cycle Deformation	1075
37.4.3	Elastic Block Modeling	1077
37.5	Seismology	1078
37.5.1	Static Displacements	1079
37.5.2	Dynamic Displacements from Kinematic GNSS	1082
37.5.3	Real-Time Application to Earthquake Warning and Tsunami Warning	1083
37.5.4	Transient Slip	1085
37.5.5	Postseismic Deformation	1085
37.6	Volcano Deformation	1088
37.7	Surface Loading Deformation	1091
37.7.1	Computing Loading Displacements	1091
37.7.2	Examples of Loading Displacements in GNSS Studies	1092
37.7.3	Loads and Load Models	1093
37.7.4	Impacts of Loading Variations on Reference Frame	1094
37.7.5	Glacial Isostatic Adjustment (GIA)	1095
37.8	The Multi-GNSS Future	1099
	References	1100

37.1 GNSS for Geodynamics

The use of geodetic positioning to study tectonic movements began nearly a century before the global navigation satellite system (GNSS) era (see [37.1] for a description), and most of the models at the core of tectonic geodesy were developed or proposed during the pre-GNSS era. The earliest studies were for displacements caused by large earthquakes, which were the only tectonic motions large enough to be observed at that time. Displacements measured from repeat triangulation before and after the 1906 San Francisco earthquake were the basis for the elastic rebound hypothesis [37.2, 3], and supported *Gilbert's* proposal that earthquakes were the result of a sudden release of strain that was built up over a long period of time [37.4]. This century-old finding remains at the core of our understanding of the relation between tectonics and earthquakes.

The earliest GNSS studies in the 1980s were carried out in an environment in which modern space geodesy had just demonstrated its full range of capabilities, although not yet its present level of accuracy. The computational tools needed to apply surface deformation to tectonic, volcanic, and other studies already existed. All that was needed was a measurement tool that could provide spatially and temporally dense measurements of surface deformation. That was provided by the first GNSS system, the global positioning system (GPS). This chapter describes how a variety of geodynamic problems can be studied using GNSS-derived displacements and velocities, and introduces the key interpretive concepts and models.

Today, data from thousands of GNSS stations are used by researchers around the world (Fig. 37.1). High-accuracy satellite orbits are available from the International GNSS Service (IGS, Chap. 33), and the stations that contribute to the IGS tracking network form a reference network that any researcher can use to determine station positions in the International Terrestrial Reference Frame (ITRF) (Chap. 36; [37.5]). Modern studies on tectonic geodesy are based on displacements or velocities derived from time series of site positions expressed in the ITRF, on regional to global scales.

The numerical tools used to relate motion and deformation of points on the surface to plate motions, fault slip, volcanic inflation, and changing surface loads are not specific to GNSS. In that sense, the specific measurement tool is not unique, and displacements or velocities measured by very long baseline interferometry (VLBI), satellite laser ranging (SLR), GNSS or terrestrial techniques can be used interchangeably as long as measurement uncertainties are well known. However, GNSS is the critical space geodetic technique for geodynamic studies because the instrumentation is

portable, inexpensive, and easy to operate, allowing GNSS networks to be deployed at a scope and on a scale that would be impossible for other geodetic techniques. Both spatial density of measurements and measurement accuracy are critical for the application of GNSS to geodynamics.

Geodynamic studies using GNSS can be grouped into several broad categories, each detailed in a section of this chapter:

- Rigid plate motion (Sect. 37.3)
- Plate boundary deformation (Sect. 37.4)
- Earthquakes and seismology (Sect. 37.5)
- Volcano deformation (Sect. 37.6)
- Surface loading deformation including glacial isostatic adjustment (GIA) (Sect. 37.7).

In each case, the development of GNSS over the last three decades has revolutionized our view of these topics. Models that allow us to relate these geodetic measurements to the causes of deformation are just as important as the measurements themselves. While mathematical derivations of the deformation models are beyond the scope of this handbook, this chapter provides references and recommendations for further reading on each topic. Derivations and detailed exploration of many of the physical models used here are provided in [37.6].

37.1.1 Accuracy Requirements

GNSS accuracy requirements for geodynamic studies depend on the expected magnitude of the displacements or rates of motion, so the types of geodynamic problems addressed by the community have expanded as the accuracy of GNSS measurements has improved. Study of the displacements caused by large earthquakes (displacements of decimeters to meters) long predated the development of GNSS, and measurement of the relative motions of the major tectonic plates (velocities of centimeters per year) was within the reach of 1980s VLBI and SLR, and the earliest GNSS studies. Many current studies focus on problems involving displacements of millimeters or velocities of millimeters per year, which were too small to measure accurately early in the GNSS era. We can expect that future improvements in GNSS accuracy will continue to open up new doors for studying geodynamic phenomena.

For many geodynamic problems, relative position accuracy is the most relevant requirement. Consider the case of a single isolated and very long fault (a fault is a fracture surface within the Earth that experiences relative motion across it; faults are often approximately

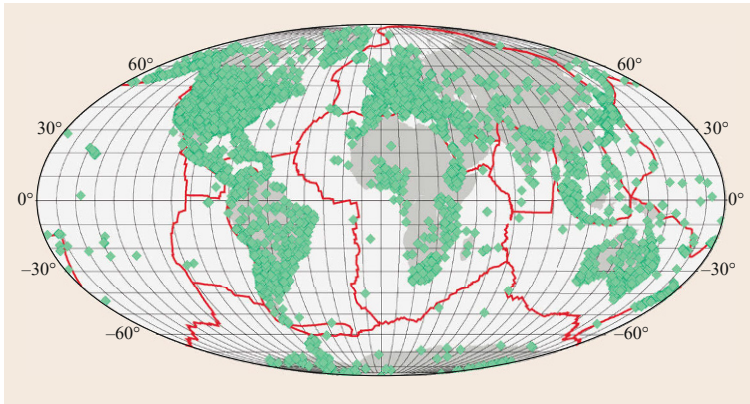


Fig. 37.1 A compilation of global continuous GNSS sites, as of 2014. *Diamonds* indicate locations of continuous GNSS sites for which data are nominally available (some countries restrict access to data from their GPS sites for legal reasons)

planar over local scales). What is most relevant for this problem is the motion of one side of the fault relative to the other. It is this relative motion that creates geologic structure and offsets, and the displacement or slip rate of the fault can be derived from a series of relative position measurements. The absolute position or motion in a geocentric frame is not strictly needed to address this problem, which is why significant advances were made in the pre-GNSS era. The accuracy required depends on the rate of motion across the fault. Ideally, the relative velocities of GNSS sites (within the local network) should be determined with a precision equal to a few percent of the fault slip rate, but useful information can still be gleaned from measurements with a lower signal-to-noise ratio (SNR).

When larger regions are considered, then the importance of a stable geocentric reference frame increases. All tectonic motions (at any scale) can be described by rotations about a geocentric axis; locally, they may be approximated in terms of linear velocities. For such studies, a stable external reference frame such as ITRF makes it possible to separate local-scale rotations and motions from large-scale rotations. For tectonic studies, displacements or velocities and their uncertainties are frequently displayed in a plate-fixed frame derived by subtracting a rigid body rotation from ITRF so that they represent motions relative to a certain stable tectonic plate.

37.1.2 Today's GNSS Accuracy

The distinction between the accuracy of relative positions/velocities and absolute geocentric positions/velocities is important because the accuracy of GNSS in these two modes is quite different. We can measure relative motions much more precisely than we can relate them to a geocentric reference frame. Analysis of continuous GNSS sites within tens or hundreds of meters of each other using a baseline solution approach

has demonstrated that relative velocities between two GNSS antennas can be measured to within a small fraction of 1 mm/yr, potentially ≈ 0.1 mm/yr [37.7]. Bennett et al. [37.8] argued that relative velocity accuracy at the submillimeter per year level could be achieved over regions of several 100 km in size. The limiting potential velocity accuracy at larger scales and absolute velocity accuracy depend on two factors: the intrinsic accuracy and long-term stability of the ITRF, and the error in accessing the ITRF (the error for the user to express their solution in the ITRF). The first error is presently on the order of 0.5 mm/yr for velocities, and is discussed in more detail in Chap. 36. One of the principal objectives of the International Association of Geodesy's Global Geodetic Observing System (GGOS) project is to improve the accuracy and long-term stability of the ITRF to the level of 1 mm for positions and 0.1 mm/yr for velocities. The error in accessing the ITRF depends on the number of ITRF reference stations used by the user and the choice of reference stations, and is not easy to quantify in general. Overall, the realistic accuracy of station velocities globally probably remains at the level of ≈ 1 mm/yr.

The precision and accuracy of site velocities estimated from GNSS time series depends on the noise characteristics of the GNSS positions. Many studies have demonstrated that GNSS positioning errors are correlated in time. Thus, estimates of site velocities or displacements (and especially their uncertainties) need to account for these time correlations either through a correlated noise model or empirical variance scaling. Assuming that each day's position errors are uncorrelated, white noise leads to over-optimistic uncertainties in displacements and especially velocities. The impact on velocities is particularly important because the error spectrum includes components that are correlated over a long time period.

The most robust estimates of the continuous GNSS position noise spectrum come from studies that use

a maximum likelihood estimate (MLE) method to fit an error model to the time series. One drawback of this approach is that it is very time consuming for long time series, but *Bos et al.* [37.9] have developed a fast method that well approximates the full MLE estimation, and is not sensitive to gaps in the data. *Hackl et al.* [37.10] and *Santamaría-Gómez et al.* [37.11] developed empirical approximations to the noise model. All of these studies found that the noise in GNSS time series can be described as a combination of white noise and flicker noise, which is a power law noise with power proportional to f^{-1} , where f is frequency. Assuming measurements were sampled evenly in time (e.g., daily or weekly), if the noise in the time series was purely white noise, the velocity standard deviation would decrease approximately as $(1/T^{3/2})$, where T is the duration of the time series. The impact of the time correlations is that the actual uncertainty (standard deviation) in the velocities decreases approximately as $(1/T)$. *Santamaría-Gómez et al.* [37.11] presented empirical relations for the velocity uncertainty based on their noise model. Based on these studies, site velocities can be as precise as ≈ 0.1 mm/yr for the horizontal and ≈ 0.3 – 0.5 mm/yr for the vertical; velocity accuracy, however, depends on the accuracy of the orientation and frame origin of the ITRF reference frame, the details of how the user accesses the ITRF, and on the magnitude of any slowly varying systematic errors in the GNSS positions.

Today's GNSS accuracy is sufficient for the study of the problems emphasized in this chapter, although the present SNR ratio can be too low for small or slow signals, such as slow-moving faults or small volcanic intrusions. GNSS random errors and periodic systematic errors remain significant compared to the size of many loading signals. Transient deformation signals of all sizes have been identified over the last decade or so, and other instrumentation demonstrates that there are transients that produce deformation too small to resolve with the present GNSS accuracy level. It is likely that smaller and smaller geodynamic signals will be amenable to study as GNSS accuracy continues to improve.

Accuracy of real-time (within seconds) or near real-time (within minutes) GNSS positioning lags well behind the accuracy of post-processed positions. This is due to two main factors:

- The lower quality of real-time orbit and clock products needed for GNSS positioning.
- The difference in positioning accuracy between epoch-wise kinematic positions and static daily positions.

The gap in quality of products is closing, and that trend is likely to continue. The difference between epoch-

wise kinematic and daily positions remains large. However, the error spectrum for kinematic positions is complex due to the effects of multipath and other errors that cannot be averaged in time, and different kinematic and real-time geodynamic applications require accuracy of ground motions at different frequency bands. Some applications are sensitive to the error at periods of one to a few epochs (e.g., time-dependent coseismic slip inversions), while others are sensitive to error at a particular frequency (e.g., study of seismic surface waves). The desired sample rate depends on the intended scientific application for the data. For example, the estimation of earthquake rupture models is currently limited to frequencies below ≈ 1 Hz because of limits in our ability to know the seismic velocity structure and calculate wave seismic waveforms. However, there is considerable information contained in seismic waveforms in the range of 1–10 Hz, so GNSS recording to 10 Hz could have considerable scientific value. This will be discussed in more detail for the problem of earthquake ground motions in Sect. 37.5.2.

37.1.3 Accuracy Limitations and Error Sources

Accuracy limitations on geodynamic studies using GNSS include the accuracy and self-consistency of the ITRF and the extent to which ITRF represents a truly geocentric frame. The ITRF is discussed in detail in Chap. 36, and only a brief summary will be presented here. The ITRF is designed to be a geocentric frame with its origin at the center of mass (CM) of the Earth system (CM frame). ITRF is a secular frame by design, and is parameterized in terms of a piecewise linear model for site position as a function of time. In terms of velocities, its present level of accuracy is thought to be ≈ 0.5 mm/yr. Important limitations to the accuracy of ITRF, and thus of GNSS positions expressed in ITRF, include the remaining systematic errors in the various space geodetic techniques, the number of co-location sites between multiple space geodetic techniques, the small number and limited accuracy of survey ties between techniques at the co-location stations, the observing site geometry of some other techniques (notably SLR), and unmodeled position offsets in the time series due to earthquakes or equipment changes. Earthquakes that occur after the definition of a given ITRF may cause displacements or nonlinear motions at certain sites, which make the ITRF positions for those sites inaccurate after such events.

The biggest question about the ITRF for geodynamic studies is how well the frame origin accurately reflects the geocenter. The frame origin of ITRF on

secular timescales is intended to be the CM of the entire Earth system, including fluids, and is thus termed a CM frame. Some geophysical models are computed in a frame with origin at the center of mass of the solid Earth (CE). The geometric center of the Earth can also be defined in terms of the center of figure of the Earth surface (CF). See [37.12] for a full description of different frames. The centroid of a sufficiently dense geodetic network should be a reasonable approximation of CF, which makes it relatively simple to define a CF frame on any timescale. Rigorous definition of a CM frame on nonsecular timescales requires detailed knowledge of the redistribution of mass within the Earth and on its surface. *Dong et al.* [37.12] argued that the current ITRF frame origin reflects CM on secular timescales, but CF on seasonal or shorter timescales. The ITRF is realized using geodetic observations and datum constraints, and thus might deviate from an ideal secular CM frame due to measurement errors, incorrect assumptions in the datum constraints, or weakness in the observing geometry of SLR (SLR defines the frame origin).

There is no universal agreement about the magnitude of any frame origin error or its cause, but several studies have made estimates of or estimated bounds on the size of potential biases. These studies agree that the largest component of any bias is likely to be in the T_z translation rate (along the spin axis). Studies using GNSS data and an assumption of rigid motions of plate interiors estimate the size of the T_z frame origin bias (rate) for ITRF2008 to be in the range of 0.5–1.1 mm/yr [37.13, 14]. *Wu et al.* [37.15] estimated a smaller bound on the T_z frame origin bias, 0.5 mm/yr, using a combination of surface velocities, GRACE and ocean bottom pressure data. The sign of the estimated T_z bias is consistent across all recent studies; this suggests that the ITRF2008 frame origin may have a small negative T_z rate bias (meaning that z velocities in ITRF are too large). However, the estimated uncertainty of the T_z bias ranges from 0.1 to 1 mm/yr in different studies, so the significance of these findings remains under investigation.

Frame origin biases impact geodynamic studies in direct and indirect ways. A nonzero frame origin bias will bias estimates of plate motion and will induce apparent internal deformation of the plates. It affects vertical velocities and quantities derived from those, and uncertainty in the frame accuracy is a major limitation in the use of GNSS data for studying small signals like sea level change.

In addition, systematic errors still remain in the GNSS position time series. Some residual systematic errors result from inadequate modeling of tropospheric path delays or other models removed from the data (ionosphere, tidal loading, etc.). These residual errors vary likely on a day-to-day or week-to-week basis, and add short-term noise to the position time series. Prominent periodic systematic errors are found at the draconitic frequency of 1.04 yr^{-1} (and its overtones), which corresponds to the period needed for a repeat of the position of the Sun relative to the GPS orbital nodes, $\approx 351 \text{ d}$ [37.16]. Draconitic harmonic errors are present in current IGS GPS products and time series [37.11, 17]. Their period has led to speculation that solar radiation pressure models may be responsible for these errors, and [37.18] demonstrated that draconitic periodic errors are substantially reduced when improved solar radiation pressure models are used. *Griffiths and Ray* [37.19] showed that errors at that period also could be caused by errors in subdaily Earth orientation parameter (EOP) models, and *King and Watson* [37.20] demonstrated that multipath also can cause spurious periodic signals aliased to the same frequencies. *Amiri-Simkooei* [37.21] estimated mean amplitudes for the draconitic variations (1.4, 1.3, 2.8 mm, respectively) for north, east and up, although this estimate could be affected by leakage from the nearby annual period. *Zou et al.* [37.22] argued that the amplitude is likely to be smaller, closer to $\approx 1 \text{ mm/yr}$, based on agreement of GPS and GRACE seasonal variations. The draconitic errors have only a minor impact on the estimation of site displacements or velocities, but must be considered when interpreting seasonal periodic deformation, for example, from seasonal surface loading.

37.2 History and Establishment of GNSS Networks for Geodynamics

37.2.1 Campaign GPS Networks

It first became possible to use GPS for positioning in the early 1980s, once four or more satellites could be tracked simultaneously. The initial test phase of the GPS satellite constellation was optimized to maximize the time for which four satellites were visible over the

southwestern United States, using the first seven operational satellites. GPS measurement campaigns for geodynamic studies began soon after, with measurements in California and in Iceland beginning by 1985. Measurement campaigns had spread worldwide by the end of the 1980s, despite the still-incomplete satellite constellation. GPS campaigns were integrated into the

European WEGENER (Working Group of European Geoscientists for the Establishment of Networks for Earth-Science Research) project beginning in the late 1980s, first focused on the Eastern Mediterranean and later extended to the whole of Europe [37.23]. GPS campaigns spanning most of Europe were organized beginning in 1989 by the EUREF project for a combination of geodetic and geodynamic objectives [37.24].

Campaign networks refer to networks of survey points that are measured episodically using GNSS instruments (Fig. 37.2). Although some campaign networks covered large areas, most tend to be spatially dense and limited in total extent, with ≈ 10 –100 measurement points spaced kilometers to tens of kilometers apart. In the 1980s and 1990, it was common to occupy each site for five consecutive days or more, although with improvements in GNSS accuracy it has become more common for each survey to occupy a site for 2–3

days, or sometimes only one. Repeat surveys for such networks are commonly on a yearly or once every few years basis.

The early 1990s mark the beginning of the modern era of GPS campaigns. Today, many groups maintain processed time series of GPS/GNSS data extending back to 1991 or 1992, while most of the data from the past are no longer used. The dividing line in time is determined by the development of a global tracking network sufficient to determine satellite orbits (and clock delay parameters) worldwide. The first dense and truly global network was deployed in January–February 1991 in the GIG 1991 campaign [37.25], and this was followed in the summer of 1992 by the IGS epoch 1992 campaign (Fig. 37.3). These temporary global campaigns, which involved the cooperation of numerous research groups and agencies worldwide, demonstrated the need for a permanent global network, which quickly led the way to the development of the IGS network (Fig. 37.3c) using a similar cooperative model [37.26].

Today, campaign GNSS surveys continue to be used to make measurements at high spatial density, often as a supplement to or densification of continuous networks. The advantage of this approach is that a large number of sites can be measured over a short time and for a very low cost compared to continuous GNSS installations. The main disadvantage is that these measurements are sparse in time, which means that time-varying signals are difficult or impossible to resolve. Campaign networks can also be prone to setup errors, and because equipment used at each site often changes, there can be additional systematic biases on each survey. Nevertheless, many important scientific results have been derived from such data. Campaign measurements often provide crucial information after large earthquakes, because in addition to measurements made by scientists, land surveyors have surveyed a large number of survey markers; re-measurement of these points can provide extremely dense earthquake displacement fields [37.27, 28].

37.2.2 Continuous GNSS Networks for Geodynamics

The first regional continuous GNSS networks for geodynamics began to develop in 1990. In southern California, the permanent GPS geodynamic array (PGGA) began with the deployment of four stations starting in early 1990 and expanded to nine sites over the next few years (Fig. 37.4). A small continuous GPS network with ≈ 1000 km spacing was established across Japan at about the same time [37.29]. These networks were established with the primary goal of measur-

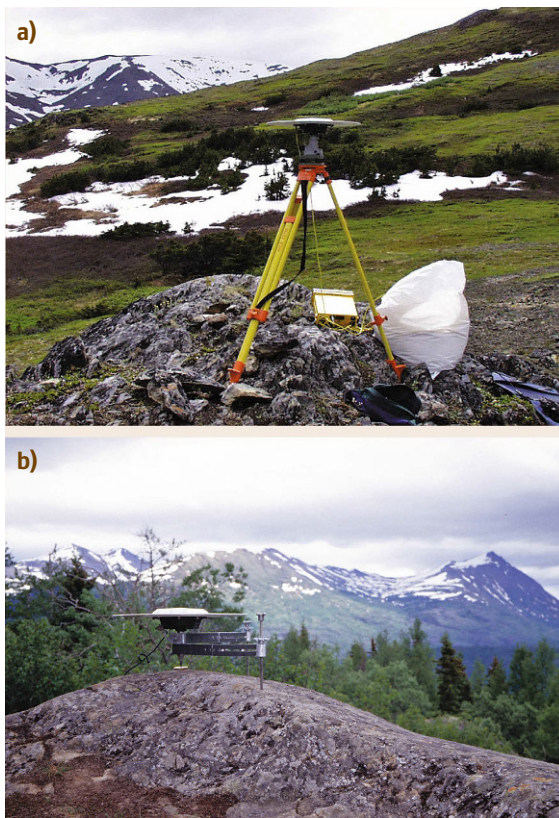


Fig. 37.2a,b Two examples of campaign GNSS sites, both from Alaska (USA). **(a)** A site setup on a tripod, with a Trimble 4000 SSE receiver and TRM22020.00+GP antenna, and a plastic bag to keep the receiver dry while deployed. **(b)** A site using the same type of antenna setup on a spike mount, a ≈ 13 cm high centering device (courtesy of Jeff Freymueller)

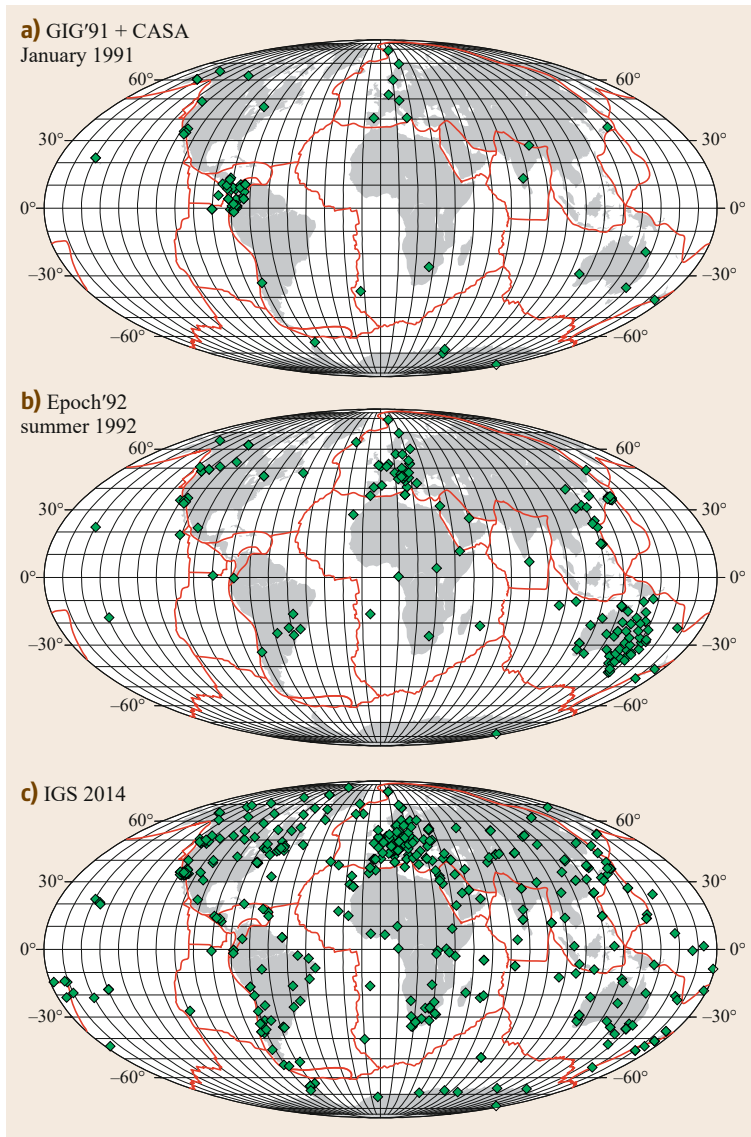


Fig. 37.3a–c Development of the global tracking network, with sites indicated by *diamonds*. **(a)** Temporary network setup for the GIG+CASA91 campaign, January–February 1991, **(b)** temporary global network for the IGS Epoch 1992 campaign, summer 1992, **(c)** IGS network as of 2014

ing deformation related to tectonic processes, rather than for purely geodetic objectives such as mapping, datum, or reference frame definition. Shortly afterward, in 1993, the BIFROST (Baseline Inferences from Fennoscandian Rebound Observations, Sealevel, and Tectonics) project established a continuous network in Scandinavia for studying GIA [37.30, 31]. These early networks successfully demonstrated the possibility and utility of measuring deformation on a daily basis. Although it may seem remarkable today, at the time the big question was whether it would be possible to process so much data on a daily basis!

The June 30, 1992 Landers earthquake in California demonstrated the power of continuous observations.

Earthquake displacements were determined from the continuous data within several days after the earthquake, rapid work for the time, and the continuous stations provided some of the first precise records of postseismic transient deformation following a large earthquake [37.32, 33]. The existence of the continuous network also provided a critical backbone for campaign studies of the coseismic and postseismic deformation [37.34, 35]. This prompted additional densification of the continuous GNSS network in both southern and northern California.

Two damaging earthquakes in 1994 and 1995 provided the impetus for a dramatic expansion of continuous GNSS networks for geodynamic studies (Figs. 37.4

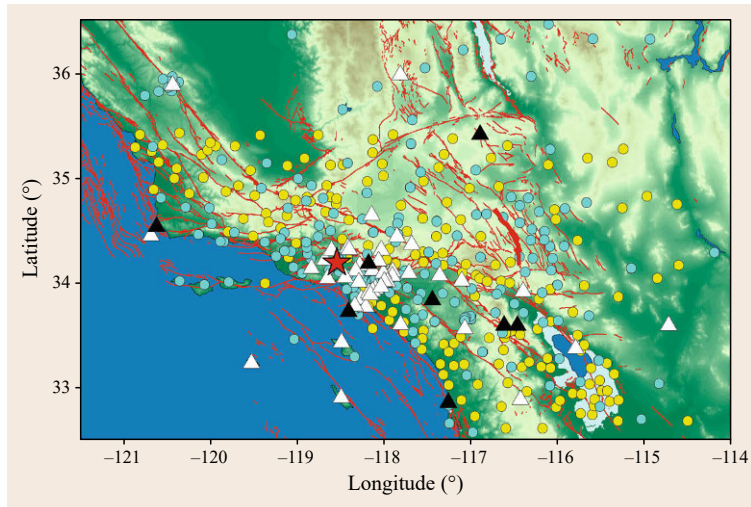


Fig. 37.4 Development of the SCIGN network in southern California, USA. *Black triangles* are the PGGA sites as of January 1994, the time of the Northridge earthquake. *White triangles* show the additional sites added after that event. *Yellow circles* show the additional sites of the SCIGN network (built 1997–2002), and *blue circles* show sites added after completion of the SCIGN network. *Thin red lines* are active faults. The *thick red line* is the rupture zone of the 1992 Landers earthquake, and the *red star* shows the location of the Northridge earthquake

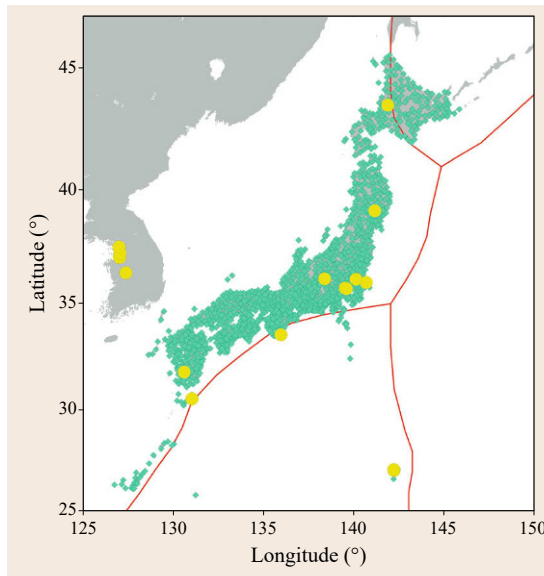


Fig. 37.5 The GEONET GNSS network in Japan. GEONET sites are shown by *green diamonds*. Sites from Japan and Korea that are contributed to the IGS network (Fig. 37.3c) are shown by *large yellow circles*. *Red lines* show a simplified view of the major plate boundaries (Japan is a region of complex deformation)

and 37.5). The Northridge earthquake on 17 January 1994 was a magnitude 6.7 earthquake that struck suburban Los Angeles, CA, USA. It killed 57 people, injured more than 5000, and caused more than \$20 billion of economic damage. The earthquake spurred an expansion of the PGGA to ≈ 57 stations within 2 years of the earthquake and later to ≈ 250 stations (as the southern

California integrated GPS network or SCIGN) [37.36]. SCIGN was followed by the plate boundary observatory (PBO) network, which spanned the entire deforming zone of the western United States [37.36], comprising ≈ 1100 stations GNSS stations including about half of the former SCIGN network, and further densification in California as part of the California real time network (Fig. 37.4).

Exactly 1 year after Northridge, the devastating Kobe (or great Hanshin) earthquake struck in Japan. This magnitude 7.3 earthquake killed more than 6000 people and caused approximately \$100 billion (10 trillion yen) in damage. In response, the Japanese government deployed the ≈ 1000 station GEONET GNSS network [37.37] across the entire country within roughly a year (Fig. 37.5). Today there are ≈ 1200 GEONET stations in Japan operated by the Geospatial Information Authority of Japan (GSI), plus several hundreds GNSS stations operated by universities or other agencies. GEONET was revolutionary in that it provided for the first time dense continuous GNSS data spanning an entire complex plate boundary. GEONET data have revealed not only steady strain within the plate boundary zone, but also a wide variety of transient deformation signals, seasonal deformation due to surface loading, and deformation due to a large number of earthquakes, including the magnitude 9.0 Great East Japan earthquake of March 2011.

Similar networks now exist across much of the world, amounting to many thousands of GNSS stations. As of January 2014, processed time series for more than 12 000 stations where, for example, provided by the Nevada Geodetic Laboratory of the University of Nevada Reno (Fig. 37.1).

37.2.3 The Importance of Global Networks

A cooperative global network was developed in parallel to these regional networks under the auspices of the IGS (Chap. 33). The IGS was originally named the *International GPS Service for Geodynamics*, and is presently called the *International GNSS Service (IGS)*. The name changes reflect both the development of other GNSS systems and the broadening of the applications of the IGS beyond geodynamic studies. The IGS network is an essential utility that enables global GNSS geodesy and is the backbone for the ITRF.

The global IGS network (Fig. 37.3c) is the glue that holds together regional networks and provides the orbit and clock products that are needed to analyze the wealth of global GNSS data. The IGS network expanded rapidly over 1992–1994 and slowly but steadily since then. Today, IGS analysis centers use as many as several hundred GNSS sites in their solutions to produce the best possible orbit and clock products, as well as a daily IGS combined network position solution that is the basis for the GNSS contribution to the ITRF. Researchers around the world use subsets of the IGS network to connect their own data to a global reference system; it functions as a global geodetic backbone to which all other data are attached.

37.3 Rigid Plate Motions

Plate tectonic theory developed in the 1960s and revolutionized our thinking about the Earth. The classic theory posits that Earth's crust is divided into a set of plates that move relative to each other, that each plate is rigid, and that all deformation occurs within narrow zones at the plate boundaries. The last of these postulates has been substantially modified with the recognition that plate boundary zones can be very broad, spanning up to ≈ 1000 km in the continents.

The motion of a rigid plate on the surface of a sphere is a rotation about a geocentric axis, and can be described by an angular velocity vector. Relative plate motions, or the motions between a pairs of plates, are also described by a rotation about a geocentric axis, and thus by an angular velocity vector. Plate motions can be estimated by determining the set of plate angular velocities in ITRF [37.5, 38], or the set of relative plate angular velocities [37.13]. The geodetic reference frame origin must be coincident with the geocenter for this description to be accurate, and [37.13] also estimated a frame origin bias in ITRF along with the plate motions.

The earliest plate models estimated from the last few million years of geologic data featured a small number of major plates (10–20). Today the number of major and minor plates is reckoned to be several times larger. Detailed models of broad deforming areas such as western North America suggest that plate tectonic-like models may apply even down to the scale of blocks with dimensions of tens of kilometers [37.39]. A further discussion of plate boundary zones appears in the next section.

Chemically, the Earth is divided into three main layers that are chemically distinct, the crust, mantle, and core. However, the boundaries in mechanical prop-

erties of the Earth do not coincide with the chemical composition boundaries. The moving plates define the lithosphere, which includes the crust and a portion of the mantle (Fig. 37.6). The plates move over the asthenosphere, which is hotter and more hydrated than the mantle lithosphere. In the simplest description, the lithosphere behaves as an elastic material while the asthenosphere is viscoelastic, which means that over very long times it behaves as a fluid. The entire mantle is viscoelastic, but the viscosity of the asthenosphere is lower than that of the deeper parts of the mantle; this allows plates to move relatively easily. The asthenosphere is likely more hydrated than the rest of the mantle, explaining its relative weakness. The existence of broad plate boundary zones indicates that the lithosphere is not perfectly elastic, but does accumulates some permanent deformation. Nevertheless, there remains a very large contrast in properties between lithosphere and asthenosphere, and over the timescales of geodetic measurements we can usually apply the simple mechanical model of elastic lithosphere over viscoelastic asthenosphere. The thickness of the lithosphere varies considerably, from ≈ 50 km within areas of active or recent tectonic deformation to > 100 km in stable continental interiors.

Given the angular velocity of a plate, the horizontal velocity \mathbf{v} of a geodetic site fixed to the crust (Fig. 37.7) is the vector cross product of the plate angular velocity $\boldsymbol{\omega}$ and the geocentric site position \mathbf{r}

$$\mathbf{v} = \boldsymbol{\omega} \times \mathbf{r}. \quad (37.1)$$

Although there are three components to the vector \mathbf{v} , when it is expressed in the local east–north–up coordinate system defined at the site, the up component

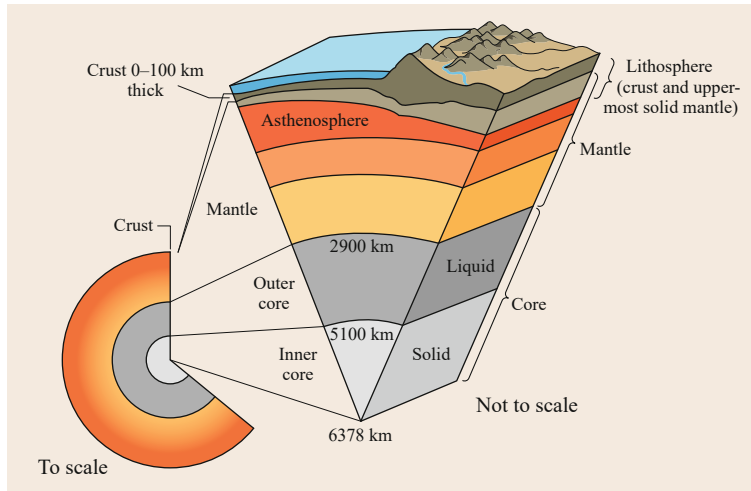


Fig. 37.6 A depth section of the Earth, showing both chemical (crust, mantle, core) and mechanical layering (lithosphere, asthenosphere, mantle). The tectonic plates are pieces of the lithosphere, which consists of the crust and the coldest part of the uppermost mantle (courtesy of United States Geological Survey (USGS)/Wikimedia Commons)

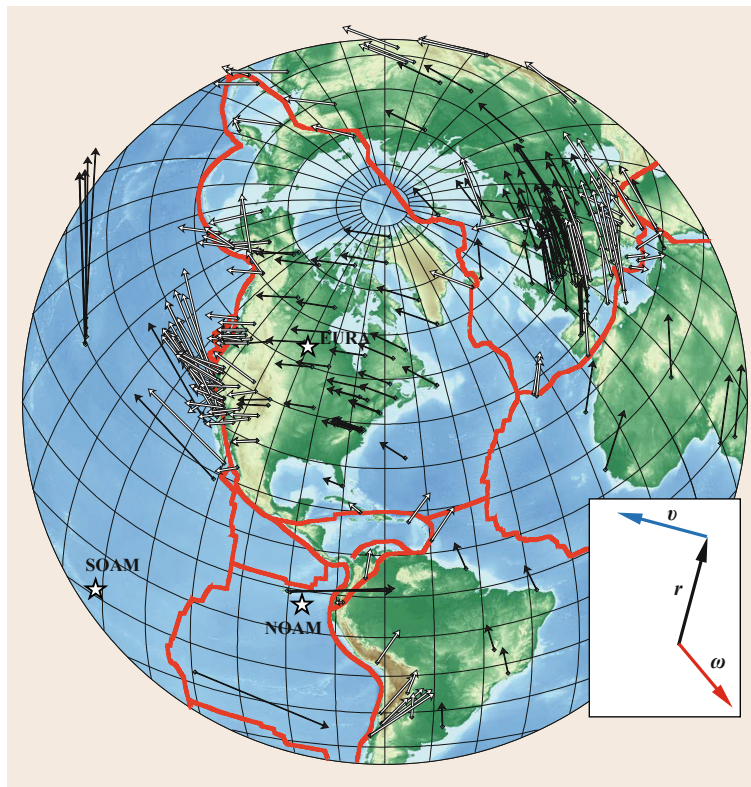


Fig. 37.7 Rigid plate motions and plate boundary zones. GNSS velocities in ITRF are from [37.13]. *Black vectors* are sites located within stable plate interiors, and *white vectors* are sites within plate boundary zones. *Stars* indicate the poles of rotation for the north American (NOAM), South American (SOAM) and Eurasian (EURA) plates. The *inset* illustrates the geometry of the cross product used to compute plate motions from the plate angular velocity. The site velocity vector is placed at the location of the site (after [37.40])

v_{up} is always zero. The simple plate tectonic model predicts that all motion is horizontal, and this approximation is good for the plate interiors, leaving aside isostatic effects. It is often convenient to describe the angular velocity in terms of a pole of rotation and angular rotation speed. The pole of rotation is the projection of the angular velocity vector onto the surface of the Earth. Given (37.1), the horizontal velocity vec-

tors v_{horiz} will rotate about the pole (Fig. 37.7). The angular speed is simply the magnitude of the angular velocity vector $|\omega|$. We can use (37.1) and the same angular velocity to describe the motion of rigid plates over geologic timescales (millions of years), and also over geodetic timescales (years). However, over geodetic timescales we also need to account for variations in deformation due to the behavior of faults

in the earthquake cycle, which will be addressed in Sect. 37.4.

It is much easier for us to visualize motions in terms of the linear velocity vector, rather than in terms of angular velocities. Over a small area, the motion of plates sometimes may be approximated by linear velocities. We can use (37.1) to evaluate the validity of this approximation, by computing the derivatives of the linear velocity \mathbf{v} with respect to the angular velocity $\boldsymbol{\omega}$ and the position vector \mathbf{r} . When the pole of rotation is far from a given point, the velocity vectors change only slowly with changes in position \mathbf{r} , and the linear velocity can provide an accurate description of plate motions over a region of hundreds of kilometers in size. However, close to the pole of rotation velocities change significantly in magnitude and orientation over small distances.

Although the mathematical relation between plate angular velocities and geodetic site velocities is very simple, there are two main sources of complexity in defining a geodetic plate motion model. First, it must be recognized which sites are representative of the stable plate interior. This task can be challenging because plate boundary zones can be very broad so that the boundary of a stable plate may be drawn incorrectly, and because some plates undergo internal deformation on short timescales due to, for example, GIA. For example, the African continent was often described as a single plate in the past, despite the extension across the East African rift system, and has since been subdivided into at least two plates. Although small, the horizontal motions due to GIA (Sect. 37.7.5) from the deglaciation after last glacial maximum (LGM) within regions that are geologically stable over the long term are significant enough to bias estimates of plate motion for the North American plate [37.41] and parts of

Eurasia. Second, the plate motion model assumes that the origin of the reference frame for velocities and the geocentric axis of plate motions is identical. If plate motions are truly horizontal, then the axis of rotation for plate motions should represent a center of figure (CF) coordinate system. ITRF is intended to represent a CM of Earth system coordinate system, and any velocity bias between the two systems will map into errors in the estimated plate motions as well as apparent internal deformation of the plates.

Dense velocity fields from GNSS can be used to test whether plate interiors are rigid. Most such studies have found that once regions of known tectonic activity or significant GIA signals have been masked out, the internal deformation of plate interiors is comparable to or smaller than the noise level in the GNSS velocities [37.13, 38, 42]. As the accuracy of GNSS velocities has improved, the upper limit on internal deformation of the plates has shrunk, with the most recent results suggesting plate rigidity to be within $\approx 0.3\text{--}0.4\text{ mm/yr}$ or smaller.

Until recently, geodetic estimates of plate motions [37.38, 43] could not distinguish any significant difference between plate motions today (measured with geodesy) and plate motions averaged over recent geological timescales (a few million years), for most plate pairs. However, Argus et al. [37.13] showed that the differences between geodetic and geologic estimates of plate motions are statistically significant given the high precision of today's measurements. This implies that changes in plate motions of a few to several percent have occurred over the last 3 million years, most notably the reduction in angular speed of the Nazca plate relative to South America [37.44]. Nevertheless, plate motions over the last few decades are similar to geologic estimates of plate motions.

37.4 Plate Boundary Deformation and the Earthquake Cycle

Although the original hypothesis of plate tectonics held that all plates were rigid and nondeforming, it was recognized from an early stage that in some areas seismicity and active tectonics were distributed across large areas, rather than being confined to narrow plate boundaries. These broad deforming regions are termed *plate boundary zones*, and are particularly common for plate boundaries involving continental crust [37.45]. Western North America and Eastern Asia are two examples of broad plate boundary zones (Fig. 37.8). Many active faults take up motion within these plate boundary zones.

The idea of an earthquake cycle is the key concept for relating geodesy to tectonics and earthquakes

(see [37.47] for a full description). This concept derives from the hypotheses of Reid and Gilbert, combined with a modern understanding of plate tectonics and the frictional behavior of faults. Earthquake cycle models describe how a steady driving stress on a fault leads to a buildup of stress and strain, which will eventually result in sudden slip on the fault in an earthquake (Fig. 37.9). A basic premise of most earthquake cycle models is the hypothesis that the parts of the fault that slip during earthquakes are frictionally locked during most of the time between earthquakes. The shallow part of a fault may not slip at all for decades or centuries, and then slip a large amount suddenly in an earthquake.

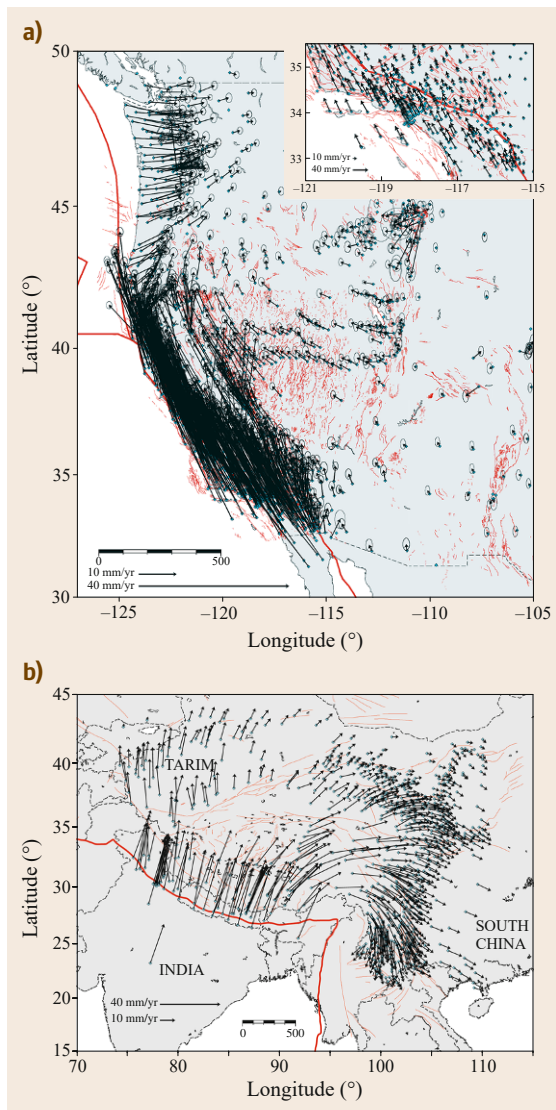


Fig. 37.8a,b Two examples of plate boundary deformation zones. **(a)** North America, showing velocities relative to the North American plate, from the University NAVSTAR Consortium (UNAVCO) PBO velocity solution. The *inset* shows details of the velocity field in southern California. **(b)** India–Eurasia, showing velocities relative to the Eurasian plate (after [37.46]). *Thin red lines* indicate active faults, and *thick red lines* show the major plate boundaries. (after [37.40])

Plate motions do not stop because shallow locked faults remain locked, so the variations in fault slip over the earthquake cycle result in deformation in the surrounding material.

Earthquake cycle models suggest that there are characteristic and complementary deformation patterns

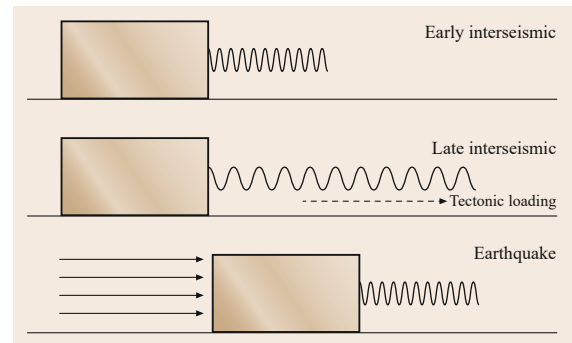


Fig. 37.9 Conceptual sketch for the earthquake cycle, based on a one-dimensional spring slider. The fault is represented by a block frictionally coupled to the surface, and the force of friction resists motion. The spring represents the elastic medium surrounding the fault. Tectonic loading steadily and slowly extends the spring, but the block does not move as long as the elastic force exerted by the spring remains smaller than the frictional force. When the elastic force exceeds the frictional force, the block accelerates because the sliding friction is lower than the static friction; this motion also compresses the spring, which reduces the elastic force and causes the block to decelerate and then stop again

between earthquakes (the interseismic phase) and during earthquakes (the coseismic phase). *Savage and Burford* [37.48] proposed a simple interseismic deformation model for a strike slip fault, and *Savage* [37.49] proposed a similar model for interseismic deformation at subduction zones. Repeated geodetic measurements after large earthquakes demonstrated the existence of a transient postseismic phase of deformation, during which deformation is distinctly different from that observed before the earthquake [37.50, 51]. These basic models have held up remarkably well over the last few decades, as the amount and quality of deformation data has exploded due to the development of GNSS. The GNSS era has seen some elaboration on these basic models, and has revealed a number of new slow and transient slip phenomena, but largely has confirmed rather than altered the basic principles of these models.

37.4.1 Plate Boundary Zones

In a plate boundary zone between two plates, the cumulative deformation across the entire zone corresponds to the total relative plate motion. This deformation is generally accommodated by a slip on a network of active faults, and often there are large nondeforming blocks surrounded by regions of more intense deformation (the term *block* is used in several ways in the geological literature, but in this case it simply means

a piece of lithosphere that is rigid and nondeforming). These features most likely develop because of pre-existing variations in the strength of the lithosphere, such as variations in rock types, pre-existing faults from an earlier phase of deformation, and so on (the continental lithosphere is extremely heterogeneous). For example, in the case of Western North America, the Sierra Nevada-great valley (SNGV) block or microplate separates the strike-slip faulting of the San Andreas fault system on its western border from the largely extensional basin and range region to its east (Fig. 37.8a). The broad plate boundary zone in Eastern Asia includes the nondeforming south China and Tarim blocks (Fig. 37.8b). Both examples shown in Fig. 37.8 involve at least three major plates, and multiple smaller blocks whose long-term motion can be described in a similar way to that of rigid plate motions.

Ignoring deformation due to the earthquake cycle for the moment, the motion of any point within a plate boundary zone can be described using (37.1), where the angular velocity is now a function of space, $\omega(\mathbf{r})$. Over most of the Earth's surface, the rigid plate–microplate description of (37.1) seems to apply until the size of the blocks approaches the thickness of the lithosphere, a few tens of kilometers, and perhaps down to even smaller length scales. See [37.45, 52] for a full discussion from this point of view.

If the entire surface can be represented by a set of rigid blocks or plates, then the angular velocity $\omega(\mathbf{r})$ will be constant within each block or plate, and will have a discontinuity across each block boundary. Alternatively, the deformation of plate boundary zones can be described using a continuous deformation model. Haines and Holt [37.53] showed that displacements, strains, and rotations could all be described in terms of the spatially variable angular velocity $\omega(\mathbf{r})$ and can be determined uniquely from this function and appropriate boundary conditions. They represented the continuous function $\omega(\mathbf{r})$ using spline functions that pass through a set of grid points; this allows the continuous function to be approximated by a discrete set of values that can be estimated from GNSS site velocities. Their approach allows for simple integration of nongeodetic data such as seismic estimates of strain rate from time-averaged earthquake moment tensors or geologic fault slip rates.

37.4.2 Earthquake Cycle Deformation

Most permanent deformation in the shallow part of the crust occurs through slip on faults. Faults are fracture surfaces within the Earth on which the two sides have been displaced. Real faults have rough surfaces with *topography* over all length scales [37.54], but faults are often approximately planar, at least over local scales.

With the exception of faults that are nearly vertical, the fault orientation generally changes slowly with depth. This section will deal with a simple geophysical approximation of a fault, in which minor deviations in fault geometry are ignored if they do not produce observable effects in surface deformation measured with GNSS. In this simple approach, we will assume that faults cut through the entire lithosphere, so that the relative motions of the two sides of the fault are taken up entirely through slip on the fault. Over a very long time (many earthquake cycles), the total slip on the fault at all depths must be equal, and this long-term average slip rate should correspond to the rate estimated by measuring long-term offsets of geological features.

Over shorter timescales, fault slip is not uniform in time or with depth. If all faults slipped at a steady rate equal to their long-term slip rate, then there would be essentially no shallow earthquakes. Earthquakes (Sect. 37.5) result from abrupt slip on faults. Large earthquakes within the continental crust rupture only a limited depth range, from at or near the surface to a lower limit that is usually no deeper than 10–20 km. Within this depth range, the frictional contact of the two sides of the fault is sufficient to prevent fault slip for long periods of time. However, the deeper parts of the fault continue to creep or shear steadily. This results in a steady buildup of elastic stress and strain in the surrounding material. Earthquakes result when the shear stress acting on the fault surface exceeds the force of friction and causes the fault to begin slipping (Fig. 37.9). For most Earth materials at the pressure and temperature conditions of the shallow crust, the initiation of fault slip causes the coefficient of friction of the fault surface to decrease, which leads to an acceleration of slip (an earthquake) and further reduction of the coefficient of friction. However, with increasing slip the elastic driving stresses are reduced, and when the driving stress drops below the frictional force again, the fault decelerates and then stops moving, and the earthquake stops. The buildup of stress and strain over time and the earthquakes that result from it are linked together as the earthquake cycle. See [37.47] for a more complete discussion of fault friction and its evolution through time, and the earthquake cycle.

Surface deformation observable by GNSS results from all phases of the earthquake cycle. Earthquakes and their effects, the coseismic phase of the earthquake cycle, are described in Sect. 37.5. The period between earthquakes is termed the interseismic phase of the earthquake cycle, and will be described in the current section. The postseismic phase of the earthquake cycle describes transient processes that occur immediately after large earthquakes, and will be discussed in Sect. 37.5.5.

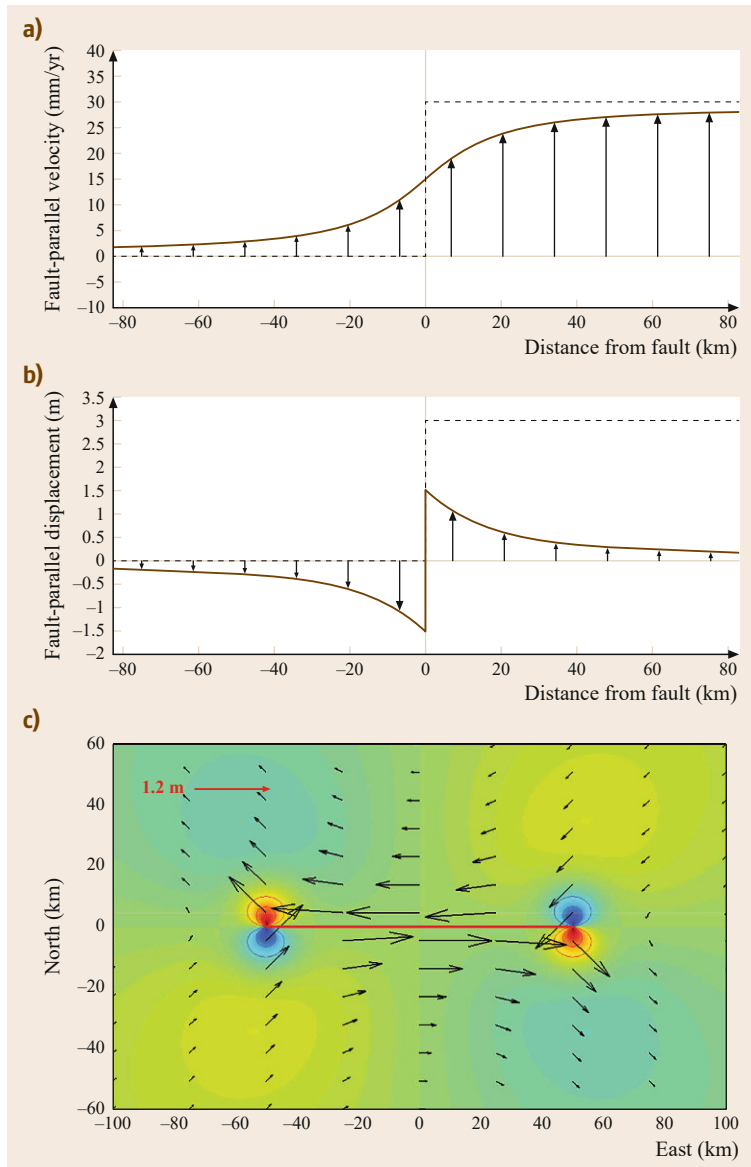


Fig. 37.10a–c Deformation surrounding a locked fault in the interseismic and coseismic phases of a simple elastic earthquake cycle. The fault in this model is a long strike slip fault (*left-lateral*) with a slip rate of 30 mm/yr, which has an earthquake after 100 years with 3 m slip. Panels a and b represent fault-normal profiles taken through the origin. **(a)** During the interseismic period, the fault is locked from the surface to 15 km depth and slipping steadily below that. The interseismic deformation pattern is shown by the *arrows* and *solid line*, relative to the far field on the left side of the fault. Compared to the uniform block motion that would result if the fault slipped uniformly at all depths (*dashed line*), the elastic response of the Earth spreads the fault shear over a broad area. **(b)** The coseismic displacement is antisymmetric about the fault, with maximum displacement at the fault (each side moves by half the coseismic slip, in opposite directions). The sum of 100 years of interseismic deformation and the coseismic displacement is a uniform block motion, shown by the *dotted line*. **(c)** A map view of the coseismic displacements from 3 m of displacement on a section of the fault. The rupture zone is shown in *red*. The colors show the vertical displacement pattern, *red* for uplift and *blue* for subsidence, with contours drawn every 5 cm (maximum vertical displacements are 21 cm)

During the interseismic period, surface deformation results from the contrast between the lack of slip at shallow depth and the steady slip at greater depth. The slip deficit is defined as the difference between the slip occurring on a part of the fault and the slip expected based on the long-term slip rate. If a part of a fault has been creeping steadily at the long-term slip rate, it will have a slip deficit of zero. Slip deficit changes with time, accumulating between earthquakes and being reduced by earthquakes. For a fully locked fault, slip deficit accumulates at the long-term slip rate. This is the model (Fig. 37.10) first proposed by *Savage and Burford* [37.48] in which the deeper part of the fault

zone creeps continuously at the long-term fault slip rate (no slip deficit), while the shallow part of the fault zone remains completely stuck by friction except in earthquakes. They represented this numerically by a planar elastic dislocation embedded in an elastic half-space that slips steadily from a locking depth d to infinite depth.

Deformation in the interseismic phase of the earthquake cycle can be computed using a superposition of steady slip on the entire fault at the long-term fault slip rate with backward slip on parts of the fault that represents the slip deficit rate [37.48, 49]. The observed surface deformation is computed by adding the surface

deformation of each of these two components together. Steady slip at the long-term faultslip rate results in the plate or block motion described by (37.1) or a local linear velocity approximation of that equation. The component due to the backward slip can be computed using elastic dislocation theory [37.55]. The elastic and viscoelastic models used to compute this deformation are explored in detail in [37.6], and their application to modeling GNSS velocity data is discussed in more detail in [37.52] and [37.40], for example. A brief description for the simplest case follows below.

For the case of a very long strike-slip fault, a two-dimensional approximation is adequate when the distance from a location to the fault is small as compared to the distance to an end of the fault. For this case, the velocity depends only on the distance from the fault. *Savage and Burford* [37.48] showed that the velocity v of a GNSS site located a distance x from the fault is

$$v = \frac{s}{\pi} \left(\arctan \frac{x - x_f}{d_1} - \arctan \frac{x - x_f}{d_2} \right), \quad (37.2)$$

where s is the long-term average slip rate, and the fault located at x_f slips from depths d_1 to d_2 . If we let $d_2 \rightarrow \infty$ and denote the upper limit of fault slip d_1 by the locking depth D , then (37.2) simplifies to

$$v = \frac{s}{\pi} \arctan \frac{x - x_f}{D}. \quad (37.3)$$

In this simple model, the fault is fully locked (no slip) from the surface to the locking depth, and is slipping at the long-term slip rate below that depth. The locking depth D thus correlates with the maximum depth of earthquake slip, usually 10–20 km for faults within the continental crust. This simple function represents the sum of the two parts of the superposition mentioned above, expressed in a reference frame in which the velocity is zero at the fault ($x = x_f$). This equation can be transformed into a version relative to the far field on one side of the fault or the other by adding or subtracting $s/2$, which is the velocity predicted by (37.3) as x goes to positive or negative infinity.

Figure 37.10 illustrates the two components of the superposition. The profiles shown in the upper two panels are fault-normal profiles through the midpoint of the fault, so that a two-dimensional approximation holds. Steady motion at the long-term slip rate would result in a velocity profile that looks like a step function across the fault. The elastic deformation due to the shallow locked fault is the difference between the solid and dashed lines, equal to $\pm s/2$ at the fault, and decaying

to zero in the far field. The locking depth controls how much of this elastic strain is concentrated close to the fault. About 50% of the elastic deformation is found within 1 locking depth from the fault, and 90% of the elastic deformation occurs within ≈ 6.3 locking depths from the fault.

37.4.3 Elastic Block Modeling

Elastic block modeling provides a way to combine the plate-like description of (37.1) and the elastic strain included in (37.3) for any network of faults, and has been applied in a wide variety of settings [37.56–60]. The model domain is broken up into a set of rigid blocks or microplates, bounded by active faults. The velocity of a GPS site is the sum of the rotation of the block it lies on and the elastic deformation caused by the slip deficit on all faults in the model. The elastic deformation is computed from three-dimensional (3-D) dislocation theory (*Okada* [37.55], for an elastic half-space) assuming backward slip on the locked part of the fault at the long-term fault slip rate. This is the same superposition described above for the two-dimensional (2-D) strike slip fault. In this approach, the fault slip rates are determined by the relative motions of the blocks on either side of the fault and the orientation of the fault. With the fault geometry specified, the motion of all GNSS sites in the model domain can be computed from a linear equation with the vector components of the block angular velocities as the parameters.

Early applications of block modeling assumed that all faults were completely locked from the surface to a constant locking depth. *McCaffrey* [37.56], with his widely-used DEFNODE software, subdivided the fault into a set of subfaults, and used a coupling coefficient that can vary across the fault so that the rate of slip deficit on each fault segment can vary between 0 and the long-term fault slip rate. This is particularly useful for subduction zones, but can be applied to any fault. *Loveless and Meade* [37.61] used a similar approach for modeling deformation in Japan, modeling the subduction zone geometry using a set of triangular dislocations. The inversion problem is linear when either the coupling coefficients or the block angular velocities are fixed; when both are estimated the inversion must be solved by a nonlinear method. Some authors have also included the option of making blocks deform internally rather than being rigid [37.60, 62]. As with the variable coupling, this extension makes the problem nonlinear except when the block angular velocities are fixed.

37.5 Seismology

Earthquakes result from abrupt and rapid slip on a fault. This induces permanent deformation in the surrounding medium (Earth), with the magnitude of the displacements scaling with both the amount of slip s and the area A that slipped. The product of these two quantities, or more properly the surface integral $\iint s dA$ over the fault surface, is called the seismic potency, and when multiplied by the elastic shear modulus of the surrounding material it becomes the seismic moment M_0 . For a point source (or small enough area A), the surface displacements everywhere are proportional to the seismic moment. Most large earthquakes involve slip over a large fault area, and for finite fault sources the displacements depend on the spatial distribution of slip on the fault.

Most people, including the general public, are familiar with the *Richter scale* of magnitudes for earthquakes. Richter's magnitude scale is technically called a *local magnitude* or M_L , because he defined the scale in terms of the amplitude of seismic waves for earthquakes in southern California, as recorded on a particular instrument (and corrected for the distance between the earthquake and seismometer). All subsequent earthquake magnitude scales have been calibrated to be consistent with Richter's original definition. Because magnitude is a logarithmic scale, an increase in magnitude by 1 represents a roughly thirtyfold increase in seismic moment. Earthquake magnitudes based on the seismic moment are called moment magnitudes, and are abbreviated as M_w , and moment magnitudes are always used for the largest earthquakes. The moment magnitude is related to the log of M_0 . For M_0 in Newton meters

$$M_w = \frac{2}{3} \log_{10}(M_0) - 6.0. \quad (37.4)$$

Earthquakes are observed to follow several empirical scaling laws. One of these shows that the average slip in an earthquake increases with increasing fault size, so that the seismic moment and thus the magnitude of displacements scale roughly with the cube of the rupture length–width. This means that large earthquakes produce much larger displacements than smaller earthquakes. These scaling relationships can change slightly for very large earthquakes because the shape of the area that slips is constrained by the physical properties of the fault. In particular, fast seismic rupture only occurs (for the most part) at relatively shallow depths, which means that rupture areas for large earthquake are usually much longer than they are wide.

The process of slip on the fault, including its acceleration and deceleration, causes the radiation of oscillatory seismic waves away from the fault surface in addition to the permanent displacements. The field of seismology focuses mainly on the observation of and interpretation/modeling of the seismic waves, which can propagate globally for earthquakes of substantial size. A wide variety of seismic waves with various propagation paths and velocities are generated by fault slip. Within the scope of this handbook, it is sufficient to divide the main types of seismic waves into body waves, which travel through the Earth, and surface waves, which travel along Earth's surface. Body waves can further be subdivided into compressional waves (P waves) and shear waves (S waves). P waves travel at a higher speed and thus arrive first, followed by S waves. Surface waves travel yet more slowly, sometimes over a longer path, and arrive even later. The Earth is mostly elastic at the timescale of seismic wave propagation, but the amplitude of the waves decreases with distance due to geometric spreading of the wavefronts and to anelastic attenuation. However, seismic waves can retain large displacements at great distances away from the rupture. Surface waves in particular can propagate long distances with large amplitudes due to their long wavelengths and 2-D geometric spreading. Unlike the seismic waves, the permanent displacements are much more localized around the region of slip.

Geodesists should view earthquakes primarily from a geometric perspective, in which motion on a surface within the Earth creates stresses that cause the surrounding medium to change its shape. In general, the pattern of static earthquake displacements can be visualized intuitively by imagining motion on a fault surface and thinking about whether the surrounding medium would be sheared, compressed, or extended by that slip. The spatial pattern of displacements from the simple case of uniform strike-slip motion on a linear fault is shown in Fig. 37.10c. Away from the ends of the fault, the displacements are parallel to the fault (which is also the direction of slip), and decrease from a maximum near the fault to zero in the far field. Near the ends of the fault, the displacements are directed toward or away from the fault, and display an antisymmetric pattern. A rough approximation of the displacements for other fault orientations can be visualized by rotating this pattern to other fault orientations; this is only approximate because the free surface (ground surface) has an impact on the displacements due to the stress-free boundary condition there.

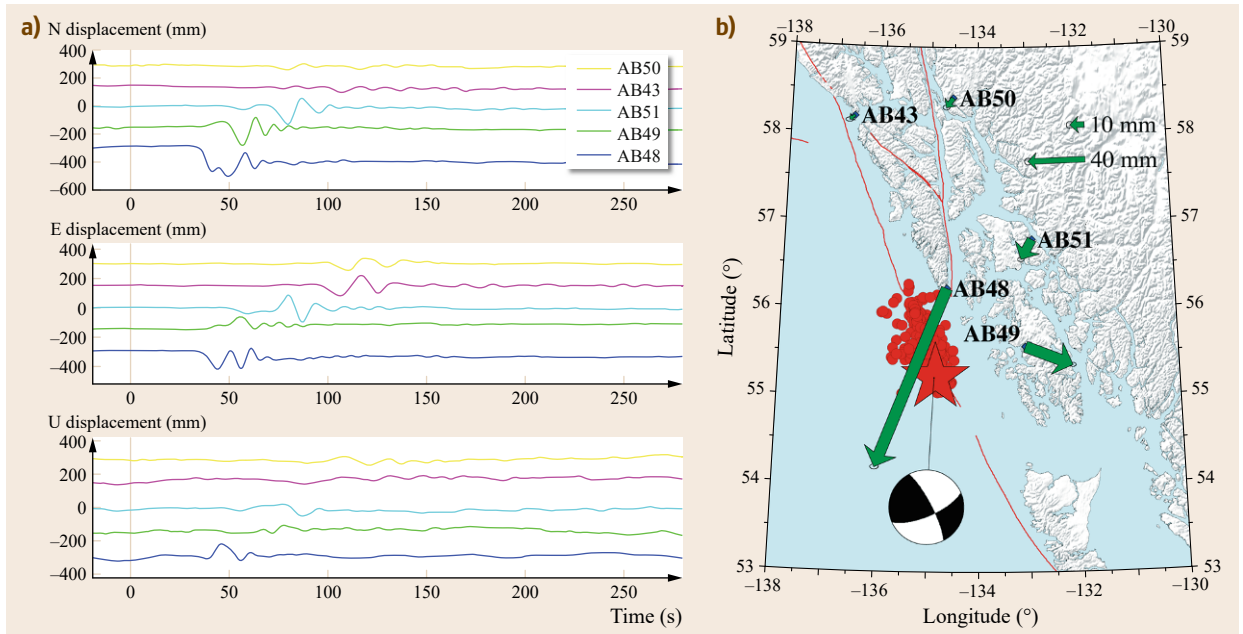


Fig. 37.11 (a) Kinematic displacement record (from 08 : 58 : 00 to 09 : 03 : 00 on 5-Jan-2013) from five plate boundary observatory sites, due to the January 5, 2013 $M_{\text{w}}7.5$ earthquake offshore Craig, Alaska. Displacements are shown for east, north, and vertical with each site having a different line style. The vertical brown line is the origin time of the earthquake. The first displacements above the noise level are seen with the arrival of the seismic S waves. (b) Map showing the site locations and the static offsets. The thin red lines are the main active faults and the cluster of aftershocks outline the earthquake rupture. The mainshock focal mechanism is also shown

37.5.1 Static Displacements

Static displacements refer to the final permanent displacements caused by fault slip in the earthquake. These displacements are not instantaneous at the time of the earthquake, but rather occur with the arrival of the seismic S waves at the site. Thus at rapid timescales (seconds) even the static displacements appear as an outward-propagating disturbance in the elastic medium of the Earth. They are recognizable as the displacement that remains once the seismic waves have propagated onward past a given site (Fig. 37.11).

Static displacements are estimated from daily or subdaily time series of GNSS positions. Averaging over several days before and after the earthquake can reduce measurement noise, but the positions after the earthquake may be changing due to postseismic deformation (Sect. 37.5.5). When displacements are small or data are limited (e.g., when only a few days of data are available), then the typical practice is to ignore the postseismic deformation and use the average position and the scatter of the daily positions to determine the offset and uncertainty. A better approach is to fit an offset for the earthquake and a time-dependent model for the postseismic deformation to the position time

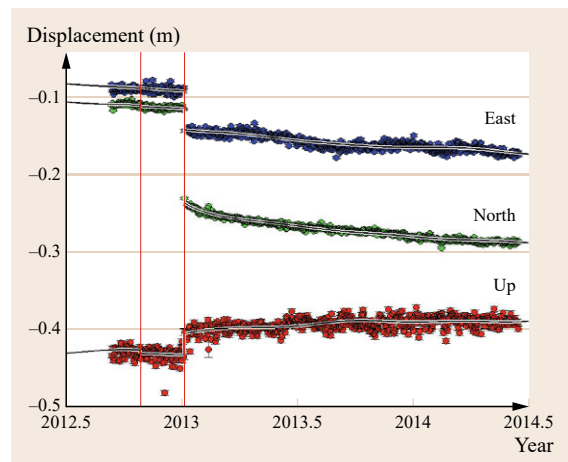


Fig. 37.12 Time series showing coseismic offset and 1.5 years of postseismic deformation for the January 5, 2013 $M_{\text{w}}7.5$ earthquake offshore Craig, Alaska. The model curve shown includes seasonal terms (estimated from the full pre-earthquake time series) and a logarithmic relaxation (Sect. 37.5.5) with a time constant of 24 d

series (Fig. 37.12). As discussed in Sect. 37.1.3, a time-correlated noise model should be used in this fit in order

to get an accurate estimate of the uncertainties of the coseismic displacement.

The pattern of static displacements is most easily visualized for the case of a long, straight strike-slip fault, that is, a vertical fault in which the two sides slip horizontally and parallel to the fault (Fig. 37.10). Direct observation of fault displacement using GNSS confirms that if the total slip on the fault is s , each face of the fault moves a distance $s/2$ (in opposite directions). For faults that break the surface, the maximum displacement usually will be observed adjacent to the fault, with the displacements decaying toward zero at greater distances. This pattern can change somewhat depending on the fault geometry and the pattern of slip; for example, if the slip at depth is much greater than the slip at the surface then the maximum displacement may be located away from the surface trace of the rupture.

Displacements as large as several meters have been observed with GNSS for several earthquakes. As most of the largest earthquakes (and thus largest slip and largest displacements) occur in subduction zones, the largest surface displacements from such events are typically offshore. However, repeat triangulation data from the great 1964 Alaska M_W 9.3 earthquake, still the second largest earthquake ever recorded, measured surface displacements as large as 20–25 m [37.63, 64]. For the 2011 M_W 9.0 Tohoku-oki earthquake, GNSS/Acoustic sites on the seafloor measured surface displacements offshore on the order of 25–30 m, which resulted from slip of as much as 50 m [37.65, 66]. GNSS/Acoustic measurements combine kinematic GNSS positioning of a ship or buoy on the sea surface with acoustic ranging to transponders on the seafloor to estimate positions of points on the seafloor.

Because the Earth behaves as an elastic body over short timescales, geodetic surface displacements from earthquakes are modeled using the elastic dislocation theory, assuming that the geometry of the fault and the distribution of slip are known. Most studies use the analytical solution for planar rectangular dislocations in a uniform elastic half-space given by Okada [37.55]. The Earth is not a uniform half-space, however, so this computation is only an approximation. Better approximations, depending on the spatial scale of the problem, can include elastic spaces with several elastic layers [37.67], or layered spherical models [37.68]. In reality, the exact distribution of slip on the fault and the fault geometry are never perfectly known. The true elastic structure of Earth is also known only approximately. Thus, it should be kept in mind that GNSS measurement accuracy for displacements exceeds the accuracy of the model calculations that link fault slip to displacements.

Displacements measured by GNSS are often used in the inverse problem, to estimate the location and geom-

etry of the fault and the magnitude of slip based on the measured displacements. This inverse problem comes in two general forms. If the geometry of the fault (location and orientation) needs to be estimated, then there is a nonlinear relationship between the model parameters and the displacements, and a nonlinear optimization approach is required. However, if the geometry is kept fixed but the distribution of slip on the fault surface is estimated, then there is a linear relationship between the model parameters (slip values) and displacements. This problem is, in general, a mixed-determined problem that can be solved with damped least squares or another generalized inverse method [37.69, 70]. It is very common to add smoothing constraints, usually the Laplacian (second derivative) of the slip distribution, into the inversion with some weight relative to the data. Positivity or other inequality constraints also act to stabilize the inversion and select against oscillatory solutions for the slip distribution. Some authors have also varied the model geometry in specific ways, and estimated a slip distribution for each model fault geometry; the best overall model is then chosen [37.28].

Figure 37.13 shows an example of coseismic displacements and a slip model, from the 2002 M_W 7.9 Denali fault earthquake in Alaska [37.27]. The inversion used 224 GPS displacement vectors plus surface offsets measured along the fault by geologists. The fault geometry was based on the mapped surface rupture, with the fault dip angle optimized where a dense GPS profile crossed the fault. The displacements at that location showed that the fault was vertical. The slip distribution was estimated using a bounded variable least squares algorithm [37.71], including smoothing via a Laplacian operator on the slip distribution. The weight given to model smoothness, relative to data misfit, is a hyper-parameter that must be chosen by the analyst. Hreinsdóttir et al. [37.27] used a combination of the model misfit/roughness *L-curve* [37.70] and a weighted cross-validation sum of squares, which tests how well a model can predict a data point that was left out in the inversion to determine the model [37.72]. Although the cross-validation approach is philosophically appealing, the displacements of sites close to the fault are so large that these sites completely dominate the cross-validation assessment. Balancing these two measures gave a range of reasonable smoothing weights. The midpoint of the range was chosen as the optimal model, but interpretations of the slip distribution were based on inspection of the entire family of models within the range of reasonable smoothing values.

Although the static displacements are localized around the area of slip, the largest earthquakes still have a very long reach given their extreme size. Six of the 13 largest earthquakes ever recorded occurred during

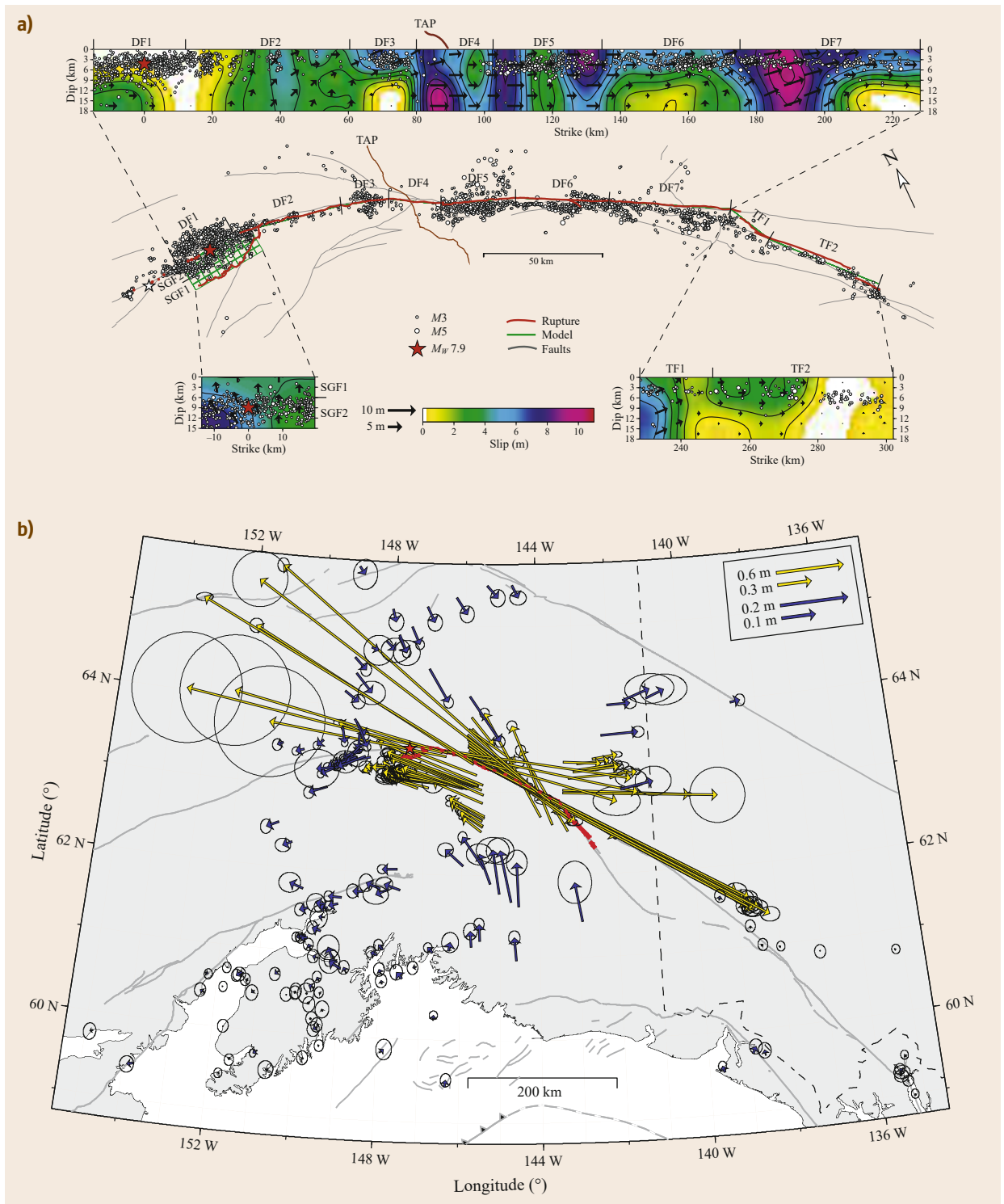


Fig. 37.13a,b Slip model and coseismic displacements for the 2002 Denali fault earthquake. The hypocentral location, the point of rupture initiation, is shown by the red star and the surface rupture shown by a thick red line. **(a)** The fault slip model with the estimated total displacement on each patch of the fault indicated by the color scale. Aftershocks are shown in map view and on the cross-section. **(b)** Map of observed horizontal displacement vectors, with two different scales used because of the ≈ 2 orders of magnitude range of displacements (after [37.27], courtesy of John Wiley and Sons)

the 8 year period 2004–2012, giving us many examples of the far-field impacts of the largest earthquakes. *Banerjee et al.* [37.73] demonstrated that the 2004 $M_W 9.2$ Sumatra–Andaman earthquake caused measurable static displacements (several millimeters) at least 4000 km away, in Korea. The 2010 $M_W 8.8$ Maule earthquake offshore of Chile caused ≈ 10 mm displacements at the Atlantic coast of Argentina, about 1000 km away. And displacements from the 2011 $M_W 9.0$ Tohoku-oki earthquake offshore of northeastern Japan could be measured across all of north China and as far away as Mongolia [37.74]. These studies show only how far away displacements can be clearly separated from the noise in the GNSS time series. The slip models for these earthquakes predict that millimeter-level displacements should extend much further away. *Tregoning et al.* [37.75] suggested that the cumulative distortion of the geodetic reference frame from the sum of these earthquakes was at nearly the millimeter level or larger worldwide.

37.5.2 Dynamic Displacements from Kinematic GNSS

Dynamic displacements refer to the seismic waves that propagate outward from the rupture. Seismic waves can be recorded globally by seismometers. The study by *Nikolaïdis et al.* [37.76] was one of the first studies to demonstrate the capability of GNSS to measure seismic waves. They processed a linked set of baselines on an epoch-by-epoch basis, and showed good agreement with long-period seismic waves. Long-period waves (such as surface waves) can have large amplitudes at great distances from the source, although shorter period body waves spread and attenuate faster. Thus, the largest dynamic displacements seen at great distances from an earthquake will most likely be surface waves. *Larson et al.* [37.77] measured the propagation of surface waves across North America from the 2002 $M_W 7.9$ Denali fault earthquake in Alaska, and demonstrated that GNSS records remained accurate and on-scale even when modern digital broadband seismometers went off-scale because of the large ground motions.

Most GNSS studies of seismic waves have used data sampled at a rate of 1 Hz. In general, this sampling rate is sufficient for estimation of earthquake source models such as finite fault rupture models. These models require accurate simulation of seismic waveforms, which are generally limited to frequencies of 1 Hz and lower by the spatial resolution of the seismic velocity structure used for the computations. However, for the 2011 L'Aquila earthquake in Italy, *Avallone et al.* [37.78] demonstrated that sampling at 5 Hz or 10 Hz was needed to provide unaliased measurements

of the displacements for sites close to the earthquake source. Important characteristics of the seismic source and the propagation medium can be determined from the waveforms directly, so for these purposes higher time resolution GNSS data would be desirable. It is not clear whether GNSS data recorded at even higher time resolution would provide important new information. GNSS positions at 5 Hz were analyzed for the 2010 $M_W 7.2$ El Mayor-Cucapah earthquake in Mexico, but the energy at frequencies higher than 1 Hz was only a minor contribution to the total signal [37.79]. The precision of high rate GNSS positions and the potential benefits of recording at a very high rate (e.g., 50–100 Hz) and filtering or averaging the resulting positions will be addressed later in this section.

Figure 37.11 shows dynamic and static displacements observed at the PBO station AB48 from a $M_W 7.5$ earthquake that occurred offshore of Craig, Alaska on January 6, 2013 [37.80]. Approximately 40 s passes from the start of the earthquake to the arrival of the first observable seismic waves at the site, located ≈ 75 km from the closest part of the rupture surface. The station position oscillates for approximately 20–25 s after the arrival of the first major displacement, eventually settling at the final static displacement. In this case, the largest static displacement is in the north component, being ≈ 10 cm to the south.

Earthquake displacement records such as this are computed through standard kinematic positioning techniques (Chap. 25). Kinematic precise point positioning (PPP) solutions are particularly straightforward for this purpose, although these solutions assume that the reference stations used for the orbit/clock solution are far enough away from the earthquake that the earthquake does not displace these stations. Otherwise, the clock and possibly orbit estimates may be biased. Kinematic baseline solutions (Chap. 26) can also provide good results, although care must be taken with baseline solutions, because these will show displacements when the seismic waves reach the *base* station in addition to the displacements of the *rover* station.

The noise spectrum of kinematic positioning is complex, and not yet fully described. Over very short times, for example, several seconds to tens of seconds, the satellite constellation geometry does not change much and propagation delays like atmospheric delays are approximately constant. Over such short timescales, the error in changes in position of a GNSS antenna is quite small, dominated mainly by the noise of the phase measurement and to a lesser extent by multipath variations. In effect, errors caused by path delays or orbit mismodeling cause systematic errors in the position but contribute only slightly to the estimate of position change. The impact of other errors depends on

the solution type; for example, interpolation of clock estimates can contribute to errors in kinematic PPP solutions. Thus, kinematic time series such as that shown in Fig. 37.11 display a variety of errors that are longer in period than the seismic waves of interest, and can be removed by filtering. *Genrich and Bock* [37.81] found that the noise spectrum of very high rate GNSS positions was nearly white for frequencies above 0.5 Hz, which means that averaging positions determined from much higher rate sampling can achieve significant noise reduction.

Because the change in position is the quantity we wish to measure to study dynamic displacements from seismic waves, the effective noise of GNSS position records is quite low and GNSS records are observed to agree extremely well with co-located seismometers, provided the seismometers remain on scale [37.77, 82, 83]. *Elósegui et al.* [37.84] tested a GPS antenna on a shake table with known motion and found that the root mean square (RMS) error of the kinematic GPS solution was quite small over 15–25 min windows. Using data from a time period in which no displacements were expected, *Genrich and Bock* [37.81] estimated that horizontal positions with a precision of 0.5 mm could be recovered by averaging positions determined at 20 Hz sampling down to a rate of 2 Hz. However, in a comparison of 5 Hz GPS records to doubly integrated accelerometer records for the 2010 $M_W 7.2$ El Mayor-Cucapah earthquake, *Zheng et al.* [37.79] estimated the error in the GPS displacements to be 4–5 mm. The precision of very high rate GNSS positions depends on the bandwidth of the phase lock loop (PLL) used within the receiver [37.85]. That study examined displacements recorded at 100 Hz sampling with PLL bandwidths of 25–100 Hz. The standard deviation of positions with a PLL bandwidth of 100 Hz was about twice as large as that for 25 Hz, and the autocorrelations of the position time series increased as the PLL bandwidth decreased. This indicates that there is minimal benefit to recording positions at a sampling rate with frequency higher than the PLL bandwidth.

Crowell et al. [37.83] proposed a method for integrating GNSS displacement records with co-located strong-motion (acceleration) seismometers. The accelerometer records must be integrated twice to get displacement, and must be corrected for any instrumental tilt caused by the earthquake displacements (static or dynamic) in addition to assuring that the integration constants are chosen properly to match the static displacements. They devised a method to combine the two data streams using a Kalman filter, and demonstrated that they could recover broadband ground motions with high fidelity. A number of continuous GNSS stations in California are now outfitted with co-located accelerom-

eters to provide enhanced data for rapid determination of earthquake magnitude and potentially earthquake early warning. *Tu et al.* [37.86] applied a similar method using a low-cost single frequency GNSS receiver and obtained ≈ 20 mm accuracy for the displacement estimates.

GNSS displacement records are now commonly used in earthquake source inversions along with seismic data. These inversions are like those described for the static displacements above, but attempt to reconstruct the time history of slip at each point on the fault and not just the total. *Miyazaki et al.* [37.82] demonstrated that such models could be derived entirely from GNSS displacement records, for the 2003 $M_W 8.3$ Tokachi-oki earthquake offshore of Hokkaido, Japan. Several more recent studies include those of the 2011 $M_W 9.0$ Tohoku-oki earthquake [37.80], the 2012 $M_W 8.6$ Wharton Basin earthquake [37.87], and the 2013 $M_W 7.5$ Craig, Alaska earthquake [37.80].

37.5.3 Real-Time Application to Earthquake Warning and Tsunami Warning

The displacements from earthquakes large enough to be damaging or deadly can be very large near the source. In particular, the displacements from large subduction zone earthquakes, which are capable of generating devastating tsunamis, can be as large as several meters in the near field. Even far-field displacements from such earthquakes can be very large compared to the precision and accuracy of real-time kinematic GNSS solutions. This motivates efforts to use real-time GNSS for earthquake early warning or tsunami warning [37.88], and systems to implement this are under development.

Real-time GNSS displacements can provide estimates of the earthquake magnitude and rupture dimensions that are highly complementary to the information available in real time from seismology. Some commonly used seismic methods to measure earthquake magnitude saturate because the seismic observations used are limited in frequency band. As a result, earthquakes much larger than $M_W 8$ can be initially reported as $M_W 8$. However, the displacements observed by GNSS do not suffer from this limitation, and the displacements from an $M_W 9$ earthquake are immediately recognizable as being much larger than those of a $M_W 8$ earthquake (Fig. 37.14). Figure 37.14 also illustrates some other properties of the earthquake source that are easily constrained by GNSS displacements. The length of the $M_W 9$ earthquake rupture is easily derived from the displacements, simply by noting where the seaward motions fall off toward zero. An experienced analyst could draw a reasonable approximation of the rupture

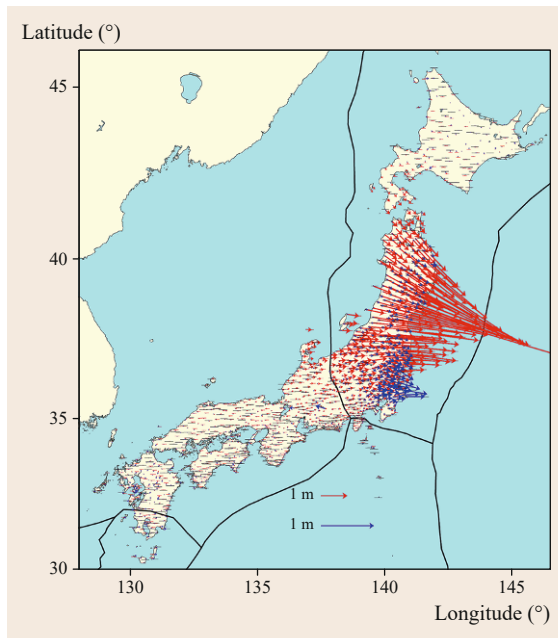


Fig. 37.14 Comparison of displacements from the March 11, 2011 M_W 9.0 Tohoku earthquake (red), and its M_W 7.9 aftershock (blue), based on GEONET data. Displacements computed by the Caltech/Jet Propulsion Laboratory (JPL) ARIA project (after [37.89]). This illustrates the vast difference in size between earthquakes 1 magnitude unit apart, and also shows how easily the rupture length can be inferred from the displacement data. Black lines show a simplified view of the major plate boundaries (Japan is a region of complex deformation)

plane from these displacements alone, and an inversion for a finite fault rupture can be done immediately once these displacements are available. The finite fault model gives information that could not be obtained from seismology until a few hours after the event.

There are several potential limitations for the application of real-time GNSS to earthquake early warning or tsunami warning. Timing and magnitude of the deformation signal are intrinsic limitations of the method, while precision of the GNSS orbit and clock products and quality control are limitations that can be addressed by improved techniques. As can be seen in Fig. 37.11, displacements of a GNSS site will be observed when the seismic body waves reach the site, not at the time the earthquake begins. This means that only GNSS sites very close to the fault rupture can contribute to earthquake early warning, which is aimed at warning of the ground shaking before it arrives. This is less of a problem for tsunami warning, as the tsunami waves travel much more slowly than the seismic waves. In addition, sites must be close enough to the earthquake

source to record significant displacements given the noise level.

Improvements in techniques will result in better orbit and clock products being available in real time. IGS ultra-rapid orbit products are already very good, because orbits can be predicted ahead by integrating the equations of motion. Clock estimates cannot be predicted ahead far in time, so real-time PPP processing is dependent on the quality of real-time estimates of satellite clock error. Current real-time clock estimates have an accuracy of ≈ 0.1 ns in the best case [37.90], which can provide positioning accuracy of 20 mm for horizontal positions and 40 mm for vertical positions [37.91]. For comparison, the IGS final clock solutions have a standard deviation half as large, while forward prediction of clock errors has an accuracy 10–20 times worse. The desired goal is to have real time clock estimates that are as accurate as the IGS final clock estimates. Probably the biggest challenge for this application is in quality control, mainly the detection of cycle slips or other data anomalies in real-time streams. In some approaches, such as that originally used by Nikolaidis et al. [37.76], ambiguities are resolved independently for each epoch, because the phase is not assumed to be continuous in time (this means there is no need to detect cycle slips). Refer to Chaps. 23, 25, and 26 for more information about the technical details of ambiguity resolution, kinematic, and differential positioning.

Use of GNSS real-time position time series for warning requires automated methods for detection and measurement of earthquake displacements, and for inversions of these data for earthquake source parameters. The details of these methods depend on whether or not it is assumed that information from seismology will be available. Rapid detection of earthquakes using seismic networks is more straightforward than using geodesy, and the tools to do it already exist. However, there have been several approaches taken to the detection problem using GNSS data alone. Ohta et al. [37.92] used a short-term average versus long-term average root-mean-square (RMS) comparison to detect both the occurrence of a displacement event and to detect when the ground motions have stabilized at the post-event positions. Even if the event detection is primarily based on seismology, it is still necessary to detect when the dynamic ground motions have finished so that the static displacements can be estimated as quickly as possible.

Real-time inversion procedures have been outlined by several authors [37.83, 92–94]. Melgar et al. [37.93] proposed using displacement time series to estimate a point source approximation for the earthquake, a CMT or centroid moment tensor solution. Crowell

et al. [37.83], Ohta et al. [37.92], and Minson et al. [37.94] outlined methods for real-time finite fault inversions. There are three basic approaches used by these authors. One approach uses a catalog of predefined faults or fault segments, which results in a linear inversion for slip on the faults. This approach may fail if the earthquake is on an unexpected fault, or involves a complex multifault rupture. Another approach uses a point source CMT solution or a set of point sources in time to define the approximate location and orientation of the fault plane, and then a nonlinear inversion or inversions for the slip distribution. The third approach uses a Bayesian inversion framework to estimate properties of the rupture plane and slip distribution. The best approach to solve this problem in general remains under investigation, but all of these authors report that reliable solutions can be obtained within a few minutes, given an accurate set of static displacements.

Two recent papers suggest that earthquake magnitude can be determined directly from the displacement waveforms, with no need for inversion. Fang et al. [37.95] showed that Gutenberg's relation between peak displacement and magnitude [37.96] holds up to nearly $M_W 9.0$. This suggests that peak amplitudes can be read directly from the displacement waveforms and used in a magnitude calculation, just as is done routinely from seismology for smaller events. Crowell et al. [37.97] identified amplitudes of P waves from combined seismo-geodetic stations, and developed scaling relationships for these as well as the peak ground displacement. These amplitudes showed a clear scaling with earthquake magnitude. These new results suggest that an initial estimation of magnitude can be made even while the ground shaking still continues.

37.5.4 Transient Slip

In addition to earthquakes, some faults also slip in a slow or transient manner. These slip events are called slow slip events, and they have been documented at most subduction zones that have adequate instrumentation. Slow slip is often accompanied by seismic tremor, and the occurrence of these together on a highly repeatable basis in the Cascadia subduction zone led to these events being termed as episodic tremor and slip (ETS) events [37.98]. Slow slip events occur in a wide range of sizes. Schwartz and Rokosky [37.99] and Ide et al. [37.100] summarized the seismic and geodetic evidence for these events; GNSS data are the critical geodetic data in most cases. Earthquakes can also trigger slow slip or creep on nearby faults [37.101, 102].

These slow and transient events challenge some of the assumptions commonly made in geodesy, such as the use of linear models to describe motion with time. Although deviations from a linear trend may be due to noise, some variations reflect real nonlinear motion. Spatial correlation of deviations from linear motion provides the key to the separation of nonlinear signal from noise. Given the quality of today's GNSS orbit and other products, position errors resulting from orbit mismodeling are small and usually will be strongly correlated over very long distances. Time-dependent deformation, on the other hand, will be correlated over much shorter distances, and will vary in space in characteristic ways that can be predicted if the likely slip region on the source fault can be identified. Modeling of slow slip events is very similar to the problem of modeling static displacements from earthquakes, apart from the different timescales involved.

37.5.5 Postseismic Deformation

Postseismic deformation refers to transient deformation following earthquakes. Large stress changes caused by large earthquakes induce changes in the pattern and rate of strain around the fault. The time evolution of strain depends on the rheology of the crust and upper mantle [37.103] and the fault geometry and coseismic slip distribution of the earthquake. Rheology refers to the mechanical properties of materials, which describe how they deform and flow. Earth materials under different temperature, pressure, and stress conditions (and different timescales) can behave as elastic, plastic/ductile, viscoelastic, or viscous materials. Postseismic deformation provides an important opportunity to estimate these properties and study the forces that drive tectonic deformation. It is thought to result from a superposition of three main physical mechanisms, which cause deformation on different spatial and temporal scales:

- Viscoelastic relaxation of the mantle and possibly lower crust
- Afterslip on the very shallow or deep parts of the fault zone
- Poroelastic relaxation.

This section focuses on the observable geodetic effects of postseismic deformation rather than the details of the mechanisms themselves. Postseismic deformation can be large and long lasting, although there is great variation from earthquake to earthquake. Postseismic displacements are always nonlinear in time. After a large earthquake, one can expect that GNSS sites over a large area surrounding it will move in a tran-

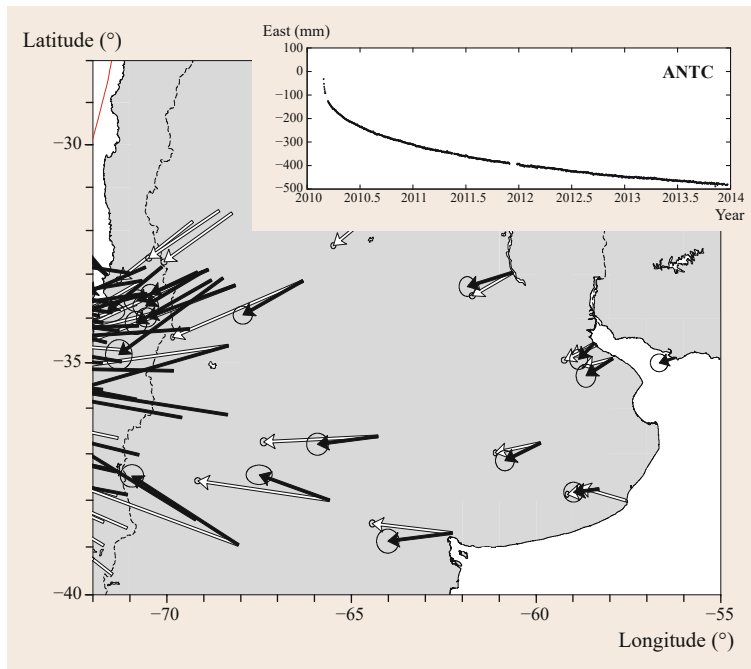


Fig. 37.15 Comparison of far-field coseismic and postseismic displacements for the 2010 Maule, Chile earthquake ($M_w 8.8$). *White vectors* are coseismic displacements (after [37.104, 105]), and *black vectors* are postseismic displacements over the first 15 months after the earthquake (after [37.105]). The *inset* shows the postseismic east component time series for the site ANTIC, which is located in Chile at the western edge of the map

sient fashion due to both afterslip on the fault zone and viscoelastic relaxation of the lower crust and/or upper mantle. This means that postseismic GNSS time series may exhibit complex time dependence. Poroeastic relaxation, which is deformation driven by fluid flow in porous media that relieves pressure differences caused by the earthquake, is generally considered to be important only very close to the fault and near geometric complexities such as fault stepovers and bends.

Afterslip

Fault friction laws and empirical models suggest that slip s on a particular spot on the fault from afterslip should follow a roughly logarithmic decay with time [37.47, 106]

$$s = S \log \left(1 + \frac{t}{\tau} \right), \quad (37.5)$$

where t is the time after the earthquake, S is a multiplicative constant, and τ is a relaxation time that depends on the frictional properties. Empirical estimates usually find a relaxation time on the order of 0.05–0.1 yr. Because the surface displacements resulting from fault slip are linearly proportional to the slip on the fault, the contribution of afterslip to GNSS time series also will follow a logarithmic decay if the spatial pattern of afterslip is time-invariant. Afterslip can occur both at shallower and deeper parts of the fault zone, compared to the slip in the earthquake. Thus, the spa-

tial pattern of displacements will generally be different from that of the earthquake. Afterslip has been identified or proposed following many earthquakes [37.32, 101, 107–111]. Theoretically, afterslip could be predicted from the coseismic slip if the fault frictional properties were known, but there appear to be spatial variations in these properties. As a result, afterslip models usually are estimated from observed displacements through an inversion similar to that for coseismic slip.

Viscoelastic Relaxation

For the simplest case of a one-dimensional material with a constant linear (Maxwell) viscosity, the displacements d from viscoelastic relaxation would follow an exponential decay with time

$$d = A \left[1 - \exp \left(-\frac{t}{\tau} \right) \right], \quad (37.6)$$

where t is the time after the earthquake, and τ is a relaxation time that depends on the ratio of the shear modulus and viscosity of the viscoelastic material. In the real Earth, viscosity varies with space (certainly with depth) and a power law rather than linear model may be expected; both of these factors will result in the exponential decay of displacements with time being an approximation only. Viscoelastic layers lie deeper within the Earth than most earthquakes, so the spatial pattern of postseismic displacements has a longer spatial wavelength than the coseismic displacements.

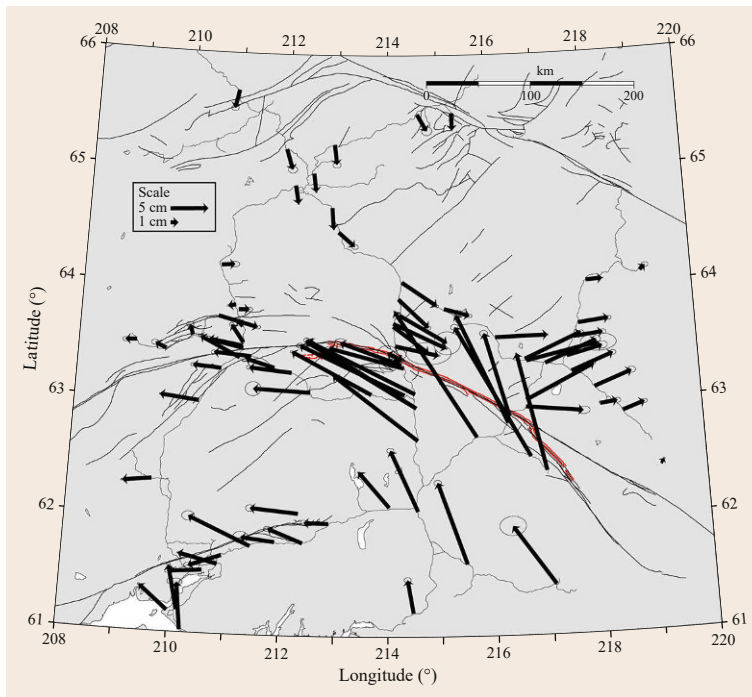


Fig. 37.16 Total postseismic displacements over a 6 year period after the 2002 Denali Fault earthquake, starting 8 months after the earthquake. The postseismic displacements were computed after removing the pre-earthquake trend from the time series, and represent the transient component of the displacement, and exceed 20 cm in places. Postseismic displacements are small close to the fault and reach a maximum at a distance of ≈ 50 km from the fault, indicating that they come from a deep source. *Black lines* are active faults, and the *thick red line* shows the extent of the 2002 surface rupture

There is considerable debate about the viscosity of materials within the Earth, but many postseismic deformation models find evidence for upper mantle (asthenosphere) viscosities in the range of 10^{18} – 10^{19} Pa s, which correspond to relaxation times of 2–20 yr, respectively [37.112]. It is important to note here that the mantle viscosity beneath tectonically active areas is lower than that below old, stable continental areas by orders of magnitude.

Examples and Implications

Adjacent to the fault, postseismic displacements are usually significantly smaller than coseismic displacements, reflecting the smaller amount of slip involved and the generally deeper deformation source. For a vertical fault, the maximum postseismic deformation is generally found 30–50 km from the fault. However, the longer spatial wavelength of the postseismic signal that results from its deeper or spatially distributed source means that far-field postseismic displacements can be a significant fraction of, comparable to, and sometimes larger than the coseismic displacements. This has been the case for several recent large earthquakes, including several in Sumatra and the 2010 Maule earthquake offshore Chile (Fig. 37.15).

Because neither the viscosity structure of the Earth nor fault frictional properties are known in detail, studies of postseismic deformation attempt to estimate them

by inversion of GNSS observed surface displacement time series. Afterslip and viscoelastic relaxation can produce similar surface deformation patterns, especially in the horizontal components, when there are enough adjustable parameters in the inversion model. This is especially true when a limited time span of data is considered, and as a result controversy over the proper postseismic deformation models can linger for years, even for well-studied earthquakes. Vertical displacements and the evolution of the transient deformation over a long time can distinguish between these physical mechanisms.

Empirical time series models are often used to describe the postseismic time evolution of GNSS coordinates. In the ITRF reference frame solutions (up through ITRF2008), only piecewise linear models are used to provide a model for coordinates, a station trajectory model, in the terminology of [37.113]. However, these models provide a poor fit to the data and/or require a very large number of parameters in cases of large postseismic displacements, such as those shown for the site ANTC in Fig. 37.15. A simple model involving one or two relaxation functions usually provides an accurate representation of the postseismic displacements over a period of several years, usually one with a short relaxation time (≈ 1 month) and one with a longer relaxation time (a few years). A complete model for the time series for one component of the position, including relaxation

terms like (37.5) and (37.6) is

$$x(t) = x_0 + v(t - t_0) + H(t - t_{eq}) \times \left[C + L \log \left[\frac{1 + (t - t_{eq})}{\tau_l} \right] + E \times \left(1 - \exp \left[-\frac{(t - t_{eq})}{\tau_e} \right] \right) \right], \quad (37.7)$$

where $x(t)$ represents the time series of some component (east, north, and up), t is time, t_0 is the reference time for the position, t_{eq} is the time of the earthquake, τ_l and τ_e are logarithmic and exponential relaxation times, respectively, and $H(t)$ is the Heaviside (step) function. The capital letters are constants to be estimated based on the time series:

- C for the coseismic displacement
- L for the logarithmic relaxation
- E for the exponential relaxation.

Bevis and Brown [37.113] showed that postseismic time series can also be usually fit well by a single logarithmic relaxation with a relaxation time of ≈ 1 yr, if data from the first few months after the earthquake are neglected.

Because of tradeoffs between parameters in curve fitting to the GNSS time series, multiple models can fit the time series well enough. This means that the time constants empirically estimated from the GNSS time series might not accurately reflect the underlying physical parameters.

The 2002 $M_w 7.9$ Denali fault earthquake provides a useful example. Ten continuous GPS sites were set up within a few weeks of the earthquake and nearly 100 sites were surveyed repeatedly over the next several years, providing a rich data set in space and time (Fig. 37.16). All sites show an initially rapid rate of deformation that decayed over time. In general, site velocities averaged over the first 2 years were ≈ 20 times faster than the pre-earthquake rates. Even several years after the earthquake, average velocities remained several times higher than the pre-earthquake rates at many sites. The time series reflects the superposition of multiple relaxation processes. No single relaxation function of time, such as an exponential or logarithmic relaxation, can explain the temporal variations of the time series to within the scatter in the measurements [37.114]. However, a combination of relaxation functions such as (37.7) can explain the temporal decay of the displacements quite well.

37.6 Volcano Deformation

Active volcanism also causes ground deformation, so surface displacements can provide critical information about the volcanic source. Injection or removal of magma in the subsurface causes changes in pressure and volume that result in surface deformation. Many deforming volcanoes have been studied using GNSS, interferometric synthetic aperture radar (InSAR), or a combination of the two. InSAR has the advantage that it can be used any time repeat satellite passes are available, and does not require access to the volcano on the ground. Remote volcanoes are thus more often studied using InSAR. However, InSAR provides only one component of displacement, which combines horizontally and vertically, so a few point displacements measured with GNSS provide an excellent complement for InSAR, and can be easier to interpret and model.

Physical sources of volcanic deformation include expansion or contraction of dikes or sills (roughly planar cracks, oriented roughly vertically or horizontally, respectively), or inflation or deflation of volumetric sources, often assumed to be spherical or ellipsoidal. Pressure changes of volumetric sources result in hor-

izontal displacements that are oriented radially away from or toward the source (Fig. 37.17) and substantial vertical displacements.

The simplest volcanic source model is the *Mogi* model (Fig. 37.18), a spherical point pressure source [37.115]. This very simple model has provided a good fit to observed displacements in many cases. It describes the deformation at the surface caused by a pressure change within a small spherical body at depth in an elastic half-space; the point source approximation actually holds even for relatively large bodies. Although other, more complex, volcanic source models exist, the Mogi source remains by far the most widely used. The success of the Mogi source is an indication that it is difficult to constrain the shape or size of a volumetric source at depth using surface observations.

The Mogi source predicts displacements that are radially symmetric about the source (Fig. 37.18). For a pressure change ΔP in a small spherical cavity of radius a

$$\Delta r = \frac{3a^3 \Delta P r}{4\mu(r^2 + d^2)^{\frac{3}{2}}}, \quad (37.8a)$$

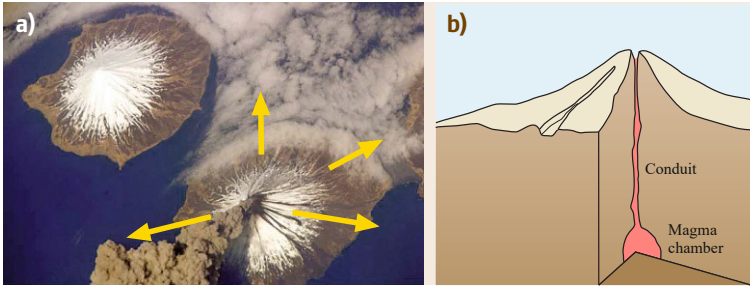


Fig. 37.17 (a) Photo of Cleveland volcano erupting in 2006, taken from the International Space Station. Horizontal displacements from volcanic sources often have radial symmetry about the volcanic vent. (b) Cutaway view of a volcano, showing a magma chamber (the most common deformation source) connected to the surface by a conduit. The base image in panel (a) is from astronaut photograph ISS013-E-24184 (courtesy of the Earth Science and Remote Sensing Unit, National Aeronautics and Space Administration (NASA) Johnson Space Center)

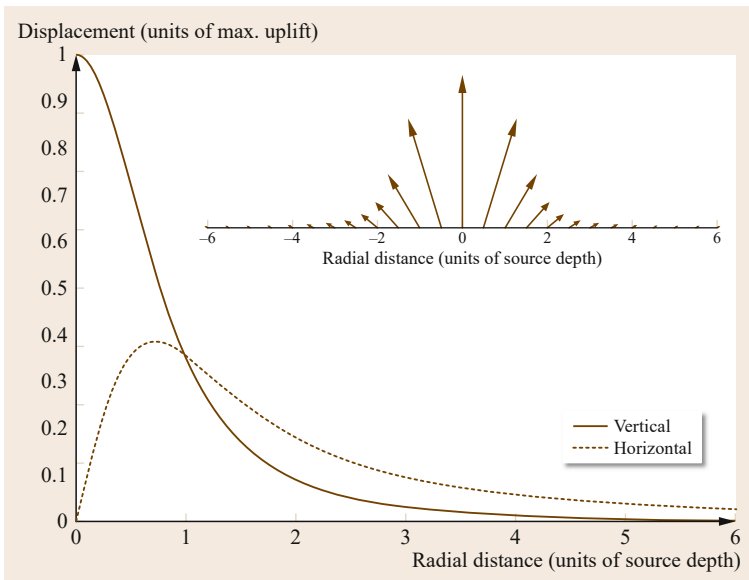


Fig. 37.18 Mogi model displacements. The blue and red curves show the vertical and horizontal displacements, respectively. The x -axis is scaled in units of source depth, while the y -axis is scaled in units of the maximum vertical displacement, which is equal to the product of the source strength and the depth of the source. The inset shows the displacement vectors along a radial cross-section

$$\Delta h = \frac{3a^3 \Delta P d}{4\mu(r^2 + d^2)^{\frac{3}{2}}}, \quad (37.8b)$$

where Δr and Δh are radial and vertical displacements, respectively, d is the depth of the source and r is the radial distance in the horizontal plane from the source to the observation point. The constant $C = 3a^3 \Delta P / 4\mu$ is called the *source strength*, and it is related to the change in cavity volume ΔV

$$\Delta V = \frac{4\pi C}{3} = \frac{\pi a^3 \Delta P}{\mu}. \quad (37.9)$$

The surface displacements in the Mogi model are linearly proportional to the volume change of the subsurface reservoir (combining (37.8a)–(37.9)). If the magma is incompressible, which should be true if it

does not contain exsolved gas bubbles, then the volume change of the subsurface reservoir is equal to the volume of magma intruded into it. However, some magmas contain exsolved gas bubbles, and this gives the magma a significant compressibility. In the presence of bubbles, the mass of the intruded magma cannot be determined from the surface deformation without assumptions about the magma compressibility. Compressible magma can accumulate in the subsurface without causing significant surface deformation, although this would still cause changes in gravity.

Many past studies have documented both inflation prior to eruption and deflation during eruption. There are a few cases of simultaneous deflation at depth and inflation close to the surface, such as the 2000 eruption of Usu volcano in Japan, for which deformation and gravity change data were fit best by a model with a de-

flation source at ≈ 4 km depth, connected to the surface by an intruded fissure [37.116, 117].

A simple *eruption cycle* model does not seem to apply to volcanoes. Unlike tectonic loading, magma supply is not uniform in time [37.118–120], and successive eruptions at the same volcano can be quite different in volume and style. It is likely that multiple bodies of crystal mush or partially solidified magma reside underneath very active volcanoes [37.121], and an eruption can be triggered when a relatively small amount of new magma rises into and remobilizes one of these bodies [37.122].

GNSS, and especially continuous GNSS (Fig. 37.19), is a powerful tool for measuring and monitoring volcanic deformation. Volcanic deformation can be quite large, involving displacements as large as meters in the case of some eruptions. Volcanoes are highly dynamic, and deformation can occur over a variety of timescales from seconds to years, such that continuous GNSS measurements are the ideal method to study them. *Fournier et al.* [37.120] provide an example of using a GNSS time series to study the long-term inflation of a volcano over several years as it was slowly built toward eruption. They used an unscented Kalman filter, a type of nonlinear Kalman filter, to estimate the volume changes over time at Okmok Volcano, revealing pulses of inflation as new magma was intruded beneath the volcano. But significant signals, especially during eruptions, can occur on very short timescales. *Larson et al.* [37.123] discussed how to optimize GNSS processing and filtering to recover deformation signals for volcano monitoring with high resolution in time. They recommended using a kinematic solution strategy that includes a degree of temporal smoothing in the GNSS solution itself, such as estimating the position using



Fig. 37.19 Photo of a continuous GNSS station for volcano monitoring, at Okmok volcano, Aleutian Islands, Alaska. The GPS antenna is mounted on a braced structure anchored in rock, with a separate hut for the receiver and batteries (*behind*). Solar panels run the instrument and its co-located seismometer. The ash deposits here are from a 2008 eruption emanating from the cone in the background (courtesy of Jeff Freymueller)

a random walk noise model with a Kalman filter. Doing so suppresses noise but still allows for accurate detection of abrupt changes in the time series (Fig. 37.20).

GNSS signals also have been used to detect volcanic ash clouds in the atmosphere. The presence of dense ash clouds can have two possible impacts on the GNSS signals and positions. First, there can be a signal path delay through the ash cloud that is large enough to bias kinematic positions, and which can

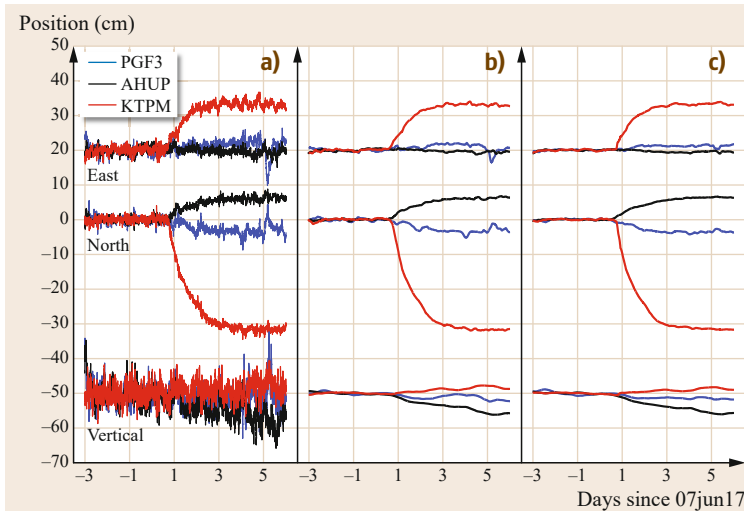
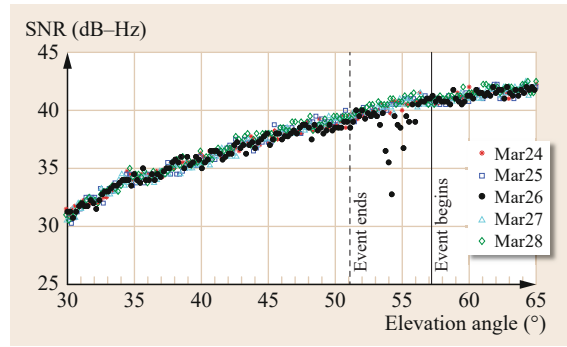


Fig. 37.20a–c Kinematic GPS position records for three stations (PGF3, AHUP, KTPM) from an intrusion of magma that preceded an eruption of Kilauea volcano, Hawaii, using three different kinematic estimation strategies (for a Kalman filter-based analysis). AHUP and PGF3 are relatively far-field stations north and south of the intrusion, while KTPM is located close to the intrusion. **(a)** Shows the positions estimated with a white noise model (independent positions each epoch), while **(b)** shows the same data but a random walk process noise model applied to the positions. **(c)** Is the same as the center panel but with gradients in the tropospheric delay also estimated (after [37.123], courtesy of John Wiley and Sons)

Fig. 37.21 Detection of volcanic ash through signal to noise ratio (SNR) variations. The *colored dots* show SNR values for the same part of the sky on several days, each shown with a different color. The eruption occurred on March 26 (*black dots*), and while the eruption plume is in place there is a period of time for which the SNR is substantially lower than normal as the signal is attenuated by passage through the plume (after [37.124], courtesy of John Wiley and Sons) ►

be detected through its impact on tropospheric path delay estimates [37.125]. The path delay most likely occurs because ash clouds can contain large amounts of water vapor. Signal paths that travel through the plume thus show substantial propagation delays relative to paths that do not. Second, the ash in the plume scatters and attenuates the GNSS signal, and can be detected through its impact on the SNR [37.124]. Larson [37.124] examined eruptions of Okmok and Redoubt volcanoes in Alaska, and showed that the ef-



fect of the plume could be seen through a temporary drop in SNR reported by the receiver (Fig. 37.21). One advantage of this approach is that no positioning solution is needed; the plume detection can be made based on the SNR variations reported by the receiver, which means it could be done in real time if data are available.

37.7 Surface Loading Deformation

This section discusses loading deformation at periods greater than daily, although the methods of computing loading deformation due to the ocean tides are identical to those used for other elastic loads. Although this section focuses on timescales longer than daily, it is essential that subdaily loading variations be correctly modeled in the GNSS processing (Sects. 2.3.5 and 25.2.3 of this handbook). Errors in modeling subdaily loading will be aliased into longer time periods, and can produce systematic periodic biases in displacements at fortnightly, semiannual, and annual periods, which might be erroneously interpreted as due to loading variations at those periods [37.126, 127].

Surface loads, such as surface water, ice, or snow, exert forces on the surface of the Earth, which cause the Earth to deform. The response of the Earth depends on the timescale for changes in loading, being elastic at short periods and viscoelastic over long periods of time. Which response is most important depends on both the spatial scale of the load and its temporal behavior. Spatially large loads cause stress changes deep within the Earth, and thus cause viscous flow more readily than spatially small loads. However, regardless of spatial scale, an elastic response is expected if the characteristic timescale of the loading variations is short compared to the viscoelastic relaxation time of the viscoelastic material. In practice, this means that loading variations at seasonal or shorter timescales can be safely treated as an elastic problem, while load changes over years to

decades might result in either an elastic or viscoelastic response, depending on the local viscosity structure. Load changes over hundreds of years or millennia, such as the load changes due to deglaciation, always require a viscoelastic model.

37.7.1 Computing Loading Displacements

Loading displacements can be computed in two general ways. For a point load or a load of a given shape and size, for example, a disk of a given radius, a Green's function can be computed to describe the surface displacements everywhere due to a load of unit size. The Green's function depends on the elastic or viscoelastic structure of the Earth and the spatial dimension of the load. More complex loads can be computed by superposition of multiple Green's functions. This approach is simplest when the load can be described by a small number of simple shapes, or if the load is given in the form of a gridded data set.

If the load is instead given as a sum of spherical harmonic basis functions, for example, a surface load model derived from GRACE spherical harmonic solutions, then the loading response can be computed easily using Love's loading theory [37.128]. Gravity changes, surface load changes, and surface displacements are all related via potential theory and elastic loading theory. The relationship between gravity and surface loads is derived from the principle of Green's

equivalent layer [37.129]. The gravity field (measured externally) can be represented mathematically as being caused by a thin layer of varying surface density at the surface of the Earth. In the case of surface loading, the surface layer refers to the time variations in the gravity field, and these are naturally visualized as arising from a thin surface layer of *load* that represents the variations in the mass of the hydrosphere and atmosphere. The same spherical harmonic coefficients (Stokes coefficients) describe the mass of the load, its contribution to the gravity field, and the surface displacements caused by the load. The spherical harmonic coefficients for surface loads derived from GRACE observations or from hydrological and atmospheric models [37.130] are thus directly related to the loading displacements.

Displacement in height can be expressed in terms of spherical harmonic coefficients for the changes in the surface load and the load Love numbers [37.128]

$$\Delta h = \frac{3R\rho_w}{\rho_e} \sum_{l,m} \frac{1}{2l+1} h_l' P_{lm}(\cos \theta) \times (\Delta C_{lm} \cos m\phi + \Delta S_{lm} \sin m\phi), \quad (37.10)$$

where R is the average Earth radius, ρ_w and ρ_e are the density of seawater (1025 kg/m^3) and the average density of Earth (5517 kg/m^3), respectively, P_{lm} are fully normalized associated Legendre functions for degree l and order m ; ΔC_{lm} and ΔS_{lm} are spherical harmonic coefficients of the variations in the surface load, and h_l' is the load Love number for degree l , and (θ, ϕ) are colatitude and longitude. In this case, fully normalized means that the spherical harmonic functions, including the cosine and sine terms, are fully normalized for integration over the sphere. The key element of this equation is that, for a load that has the spatial form of a spherical harmonic basis function, the displacement is proportional to the load, with the load Love number giving the constant of proportionality. Load Love numbers depend only on the elastic structure of the Earth and are provided by Farrell [37.131] and other authors for various Earth models. Similar equations are used for the horizontal displacements [37.132].

A more complete description of the mathematical theory for elastic and viscoelastic loading computations can be found in other sources, for example, [37.133]. The solution for the viscoelastic problem can be derived from the solution for the elastic problem using the correspondence principle. When the load is represented in spherical harmonic basis functions, the viscoelastic problem has a similar form to (37.10), except that the Love numbers are a function of time. In detail, there are a variety of computational methods used to solve viscoelastic loading problems for GIA, using different approaches to compute the deformation rapidly

and accurately. A recent benchmarking study that illustrates the variety of software developed for the problem, and assesses their computational accuracy, is given in [37.134].

37.7.2 Examples of Loading Displacements in GNSS Studies

The elastic structure of the Earth is known very well independent of geodesy, so loading models and GNSS displacements should agree with each other well. However, systematic errors in the GNSS position time series or in the loading models will result in misfit, and loading effects can bias reference frame realization, as loading deformation can be aliased into frame transformation parameters [37.28, 135, 136]. Heki [37.137, 138] demonstrated that seasonal deformation in Japan could be explained primarily in terms of snow loading. Dong et al. [37.139] assessed seasonal deformation globally and found that about 40% of the variations could be explained in terms of loading variations. Blewitt et al. [37.140] used the IGS network to detect global degree-1 deformation associated with seasonal inter-hemispheric mass transport.

Before about 2009, more detailed comparisons between loading model predictions and GNSS coordinate variations did not agree very well. Van Dam et al. [37.130] did such a comparison for Europe, and found poor agreement between observed and predicted displacements outside of a few areas with very large loading signals. They suggested that the discrepancies were mainly caused by systematic errors in the GNSS time series. GNSS time series based on systematically reprocessed data show much better consistency with the predictions of loading models, so it is now clear that these systematic errors have been reduced. Recent work in several areas of large magnitude surface loads has shown very good agreement (Fig. 37.22) between seasonal load models or GRACE observations and GNSS coordinate variations [37.132, 141–144].

Loading displacements can be extremely large. Seasonal vertical displacements in the Amazon basin have an amplitude of as much as 40 mm, ≈ 80 mm peak to peak, and are described well by loading models based on GRACE [37.145]. Even horizontal GNSS displacements there are coherent and in agreement with the loading model predictions [37.132]. Seasonal displacements in West Africa [37.142], the Himalaya [37.143], and Alaska [37.144] have amplitudes of 10–20 mm, while displacement due to river loads in Bangladesh can approach 30 mm amplitude [37.146].

Care must be taken in comparing GNSS displacements to a loading model, depending on what components are removed in the processing of the data.

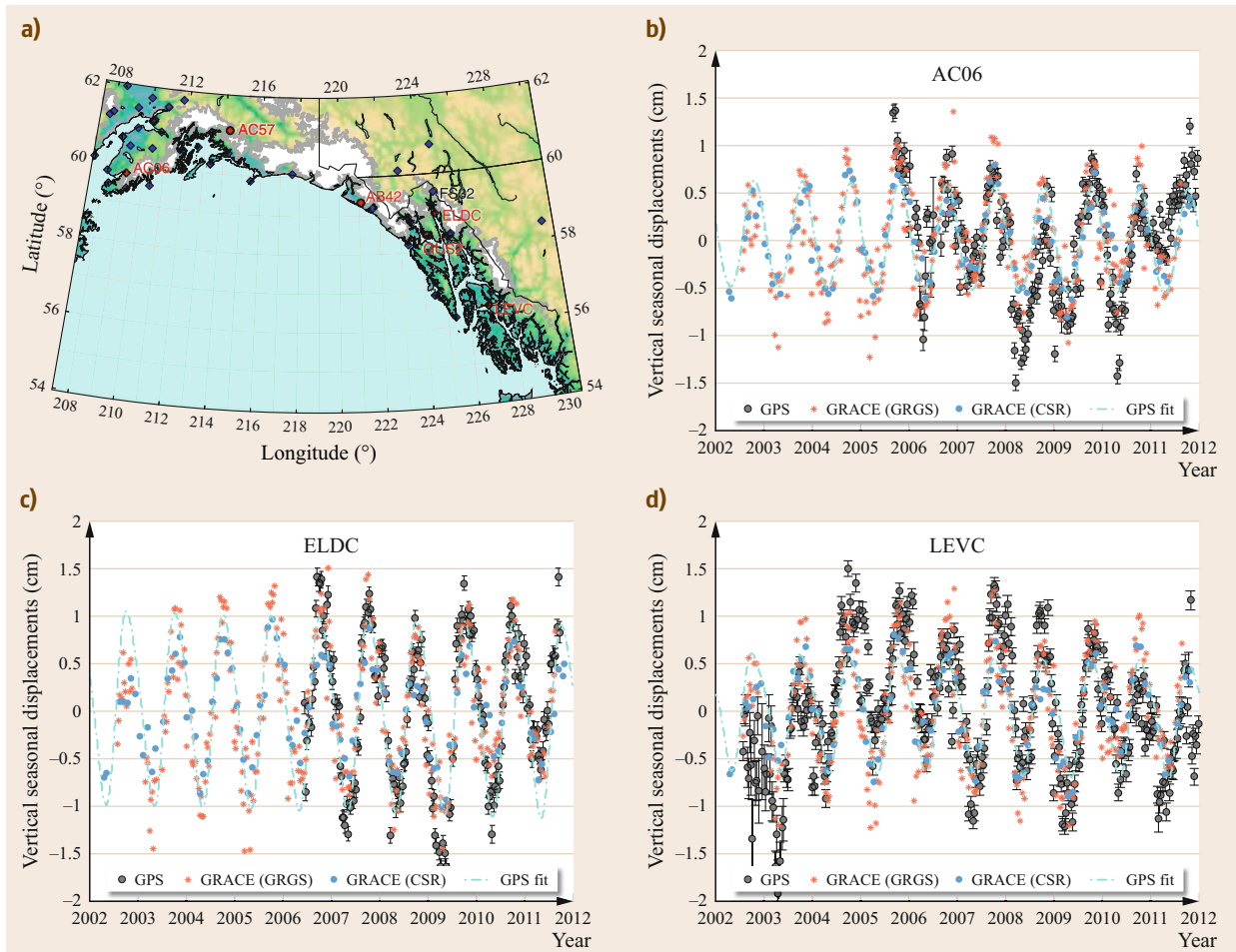


Fig. 37.22a–d Comparison of GPS observed and GRACE predicted seasonal displacements for three sites in Alaska (after [37.144]). Sites are labeled on the map in (a). (b–d) Black dots are 10-day averaged GPS height estimates, red dots are the predictions of a GRACE model based on the GRGS analysis center, and blue dots based on the CSR/University of Texas analysis center. All time series have been detrended. Different GRACE solutions show only minor differences in predicted displacements, and agree with GPS in amplitude and phase of the seasonal signal, and also in interannual variations. The AOD1B antialiasing model has been added back to the GRACE solutions, so all time series show loading due to the sum of continental hydrology, atmospheric pressure and nontidal ocean variations (after [37.144], courtesy of John Wiley and Sons)

For example, many GRACE loading models represent continental hydrology only, as atmosphere and ocean loading effects have been removed in the GRACE processing. In order to compare these models to GNSS time series, either atmospheric and ocean loading effects must be removed from the GNSS positions, or they must be added back into the GRACE model by adding the AOD1B antialiasing model. See [37.147] or [37.144] for a full discussion and examples. Removing atmospheric and ocean loading deformation from the GNSS positions is the logical approach when the goal is to use the displacements to study continental hy-

drology. Adding the AOD1B antialiasing model back to the GRACE gravity fields is the logical approach when the goal is to compare the total loading displacements between the two geodetic measurement systems.

37.7.3 Loads and Load Models

The following subsections briefly describe the models available for different loading components. Time series of load models and predicted displacements for many loading sources can be found at the Global Geophysical Fluid Center (GGFC) [37.148].

Nontidal Ocean Loading

Nontidal ocean loading refers to the loading due to nontidal oceanic mass redistribution, such as seasonal changes in freshwater runoff, sea surface height, salinity, or variations in oceanic dynamic topography due to changes in currents, winds, and so on. For sites near the coasts, this can be a significant cause of deformation. Nontidal ocean loading models are based on global ocean bottom pressure models such as the ECCO model [37.149].

Atmospheric Loading

Atmospheric pressure loading is due to spatial variations in atmospheric pressure acting on the surface of the Earth. Pressure variations can occur rapidly, changing substantially over a matter of hours, but can also be relatively stable over much longer time periods (several days) in some places. *Bevis et al.* [37.147] demonstrated in the case of Greenland that the average seasonal variation in atmospheric pressure is large enough that it contributes nearly as much to the seasonal position variations as the seasonal mass variations of the Greenland ice sheet. However, outside of the polar latitudes, atmospheric pressure loading is dominated by short-term load variations. Data for the atmospheric pressure variations usually come from operational or reanalysis products such as the NCEP reanalysis product, or equivalent from the European Centre for Medium Range Weather Forecasting (ECMWF). A method for computing the deformation from atmospheric pressure loading was proposed by *van Dam and Wahr* [37.150]. They accounted for pressure variations within ≈ 1000 km of the station, and assumed an inverted barometer response of the oceans to atmospheric pressure.

Some researchers have proposed that atmospheric pressure loading should be removed from GNSS and other geodetic time series on a routine basis because global atmospheric pressure variations are well known from global atmospheric models. Doing so will remove some short-term noise in the time series, and some of the seasonal variation in parts of the world that have a strong seasonal cycle in average atmospheric pressure. However, removing atmospheric pressure loading alone may leave large loading variations still in the position time series, because atmospheric loading usually is not the dominant component of the load model.

Continental Hydrology Loading

A variety of models for continental hydrology are now available. Development and validation of such models are an active area of research, especially given the critical new constraints on long-wavelength movements of

continental surface water that come from the GRACE mission. Some of the available models for continental water storage include the GLDAS model [37.151], WGHM (WaterGAP Global Hydrologic Model) from the University of Kassel and University of Frankfurt, and MERRA-Land from NASA Goddard [37.152]. Geodesists are actively using and evaluating all of these models. *Li et al.* [37.153] found that MERRA produced better agreement with GPS seasonal variations than GLDAS (they did not evaluate WGHM).

37.7.4 Impacts of Loading Variations on Reference Frame

Seasonal variations are not included in the ITRF, although they are present in GNSS time series. Unmodeled seasonal variations at sites used for reference frame alignment are aliased into the reference frame parameters and can bias all coordinates in the transformed solution. As a result, real seasonal variations at sites used for reference frame alignment affect coordinate time series at a seasonal timescale. Ignoring true seasonal variations will bias both the reference frame parameters and the coordinates of all sites. *Collilieux et al.* [37.135] investigated this problem extensively, using a suite of forward models and studies on synthetic data sets, mainly from the point of view of methods to mitigate the aliasing of seasonal variations into global terrestrial reference frame (TRF) parameters. They concluded that the impacts of loading variations are significant, and the reference frame aliasing effect can exceed the millimeter level in coordinates and larger than that in frame parameters.

Future reference frame models might include standard conventional models to approximate the displacement due to loading on a seasonal or other basis. If so, it is not yet clear what models would be chosen to represent these displacements. *Zou et al.* [37.22] showed that a seasonal augmentation to the ITRF is needed in order to make an accurate comparison of GPS time series and loading models, and that use of ITRF2008 without seasonal terms causes the amplitude of seasonal variations in the coordinate time series to be damped down relative to the true loading deformation. Furthermore, *Zou et al.* [37.22] compared a loading model based on GRACE to the suite of forward models used by *Collilieux et al.* [37.136]. The model based on GRACE performed substantially better, although performance over most of North America was similar to the forward models. It is likely that deficiencies in the continental hydrology models in areas of limited terrestrial data are responsible for this difference.

37.7.5 Glacial Isostatic Adjustment (GIA)

Great ice sheets covered much of the northern parts of North America and Scandinavia until the LGM, about 23 000 years ago. At that point, the ice sheets began to disintegrate due to a warming climate, and were largely gone by 8000–10 000 yr ago [37.154, 155]; the Greenland and Antarctic ice sheets remain but also have lost mass since LGM. The mass of the ice sheets depressed the surface and induced stresses that caused the mantle to flow away from the loads, even at great depths. Upon retreat of the ice sheets, the process was reversed and the mantle is still responding to the load changes caused by deglaciation, resulting in a significant GIA deformation signal. Note that the load changes include both the loss of ice in glaciers and ice sheets, and the increase of water in the oceans. In Scandinavia and the Hudson's Bay region, present-day uplift rates are on the order of 10–12 mm/yr [37.30, 41, 156]. GNSS has become an essential tool for studying the response to these past ice changes, and for quantifying present-day ice mass changes globally.

The most rapid GIA uplift rates today arise not from the areas formerly covered by the great ice sheets, but from areas with regional ice masses (glaciers and ice fields) that have been losing mass recently. Ice mass loss in Alaska and surrounding parts of Canada in North America and in Patagonia in South America has been large enough since the end of the little ice age (LIA) \approx 200 years ago to account for 10–20% of twentieth century global sea level rise over that time [37.159–162]. In both of these regions, measured uplift rates exceed 30 mm/yr at their peak, and significant uplift has been observed in a broad area surrounding the regional ice fields. In Alaska, the observed uplift depends significantly on both the ongoing regional ice mass loss and the nineteenth century deglaciation of Glacier Bay [37.163]. The former source has important elastic and viscoelastic components, while the latter is purely a viscoelastic response, because the main ice loss in Glacier Bay occurred more than a century ago.

GIA from Last Glacial Maximum

The BIFROST project [37.31, 156, 157] began continuous GPS measurements in Sweden and Norway in 1993 to study GIA, tectonics, and sea level change across the former extent of the LGM Fennoscandian ice sheet. The network later expanded to cover Finland and other countries around the Baltic Sea, and the sites eventually became part of the national networks for positioning and mapping.

Lidberg et al. [37.31] summarized the results based on 13 years of continuous GPS data, beginning with

data from mid-1996 (Fig. 37.23a). Their velocity solution had an internal consistency as good as 0.2 mm/yr for horizontal velocities. Their preferred *realistic uncertainty* estimate (based on a feature of the GAMIT software) was, typically, a factor of 2–3 larger than an estimate based on a white noise assumption, and horizontal uncertainties were commonly < 0.1 mm/yr with vertical uncertainties commonly < 0.3 mm/yr. Independent analyses of the data using GAMIT and GIPSY agreed at the 0.1–0.2 mm/yr level. They found that the data (Fig. 37.23a) could be explained by a GIA model (Fig. 37.23b) with a misfit at the level of (0.2, 0.3, 0.3 mm/yr) for the (east, north, and vertical) components, respectively, once a 0.9 mm/yr vertical bias between the observations and the model was removed (the GPS vertical velocities were larger than the model). This estimated frame bias is within the range of the values estimated later for the frame origin bias in ITRF, discussed in Sect. 37.1.3. Their optimal GIA loading model featured a 120 km thick elastic lithosphere and a two-layer viscoelastic mantle. The viscosities of the mantle layers were $5 \cdot 10^{20}$ Pa s for the upper mantle and $5 \cdot 10^{21}$ Pa s for the lower mantle, which is very similar to the VM2 and VM5a models [37.164, 165] shown in Fig. 37.24.

Compared to Scandinavia, a smaller GNSS data set is available for study of the Laurentide ice sheet in North America, at least in terms of stations located above the main loading center in northern Canada. However, there are many sites located near the southern edge of the LGM ice sheet in the United States. Sella et al. [37.41] produced a velocity field of 362 GPS sites, of which 123 were campaign sites of the Canadian Base Network with several occupations over an 11 year span, and the remainder continuous sites. Vertical velocities due to GIA in Eastern North America are, roughly speaking, uplift in Canada over the former LGM ice sheet, and subsidence in the United States over the formerly uplifted forebulge.

Although the GIA signal is much larger in the vertical component than in the horizontal (Fig. 37.23), in North America the region with significant horizontal deformation due to GIA is quite substantial. Sella et al. [37.41] used their vertical velocity field to mask out sites likely to have significant horizontal motion due to GIA so that they could estimate the angular velocity of the North American plate without bias due to the GIA horizontal deformation. Horizontal displacements from different mantle viscosity structures are quite different, but horizontal displacements are also very sensitive to lateral variations in the Earth structure [37.166]. In addition, the long length scale of the deformation can make it difficult to isolate the relatively small horizontal GIA signal from plate motions. These factors have

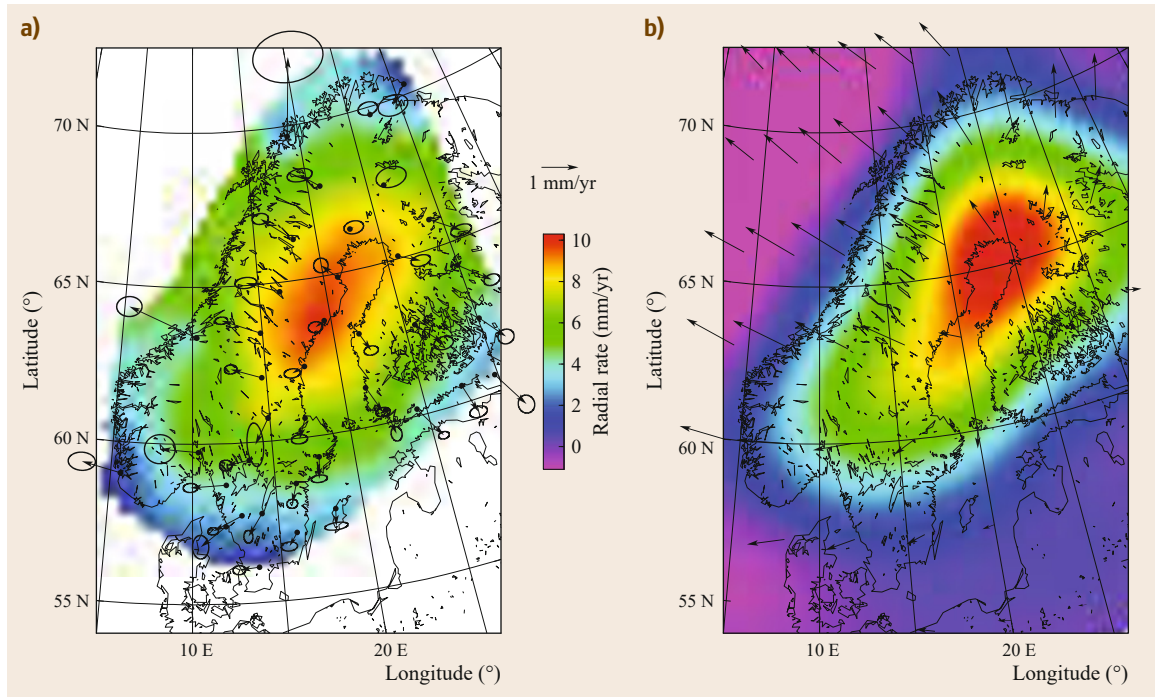


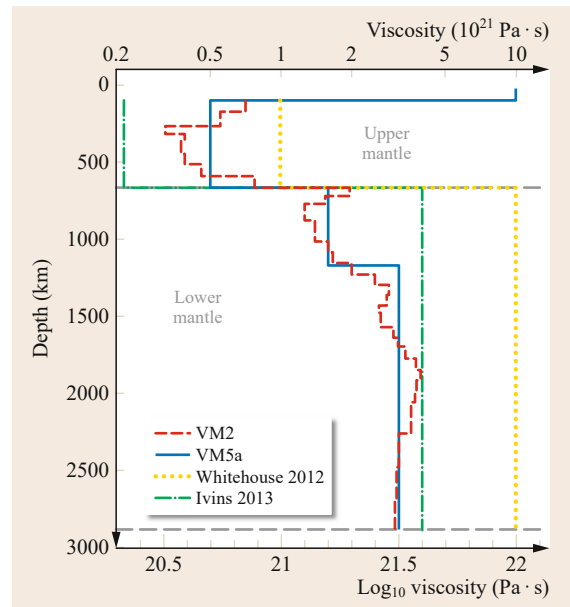
Fig. 37.23a,b Observations and model predictions for GIA for Scandinavia, using the BIFROST solution of (after [37.157]). **(a)** Horizontal velocities relative to the Eurasian plate (*arrows*, with 95% confidence *ellipses*), and vertical velocities (*background color*). **(b)** Model predictions based on the ICE-3G ice model (after [37.158]), with lithosphere thickness 120 km, upper mantle viscosity $5 \cdot 10^{20}$ Pa s, and lower mantle $8 \cdot 10^{21}$ Pa s (after [37.40])

Fig. 37.24 Viscosity structure for several mantle viscosity models. Models for low viscosity regions like Alaska, Patagonia and the Antarctic peninsula are similar at depths greater than 300 km, but require a thin (50–60 km) elastic lithosphere and an asthenospheric viscosity 1–2 orders of magnitude lower than models shown here (after [37.165], courtesy of Oxford University Press) ►

made the comparison of horizontal GIA observations and models challenging and subject to vigorous debate.

Post-Little Ice Age (LIA) GIA in Alaska and Patagonia

Glaciers in the coastal mountains of Alaska (USA) and British Columbia (Canada) began losing enormous amounts of ice approximately 200 years ago, after the end of the little ice age (LIA). The uplift and ice unloading history there are constrained by observations of glacial moraines and trimlines, raised geomorphic shorelines, tide gauge, and GPS observations [37.163]. Of particular note is the deglaciation of Glacier Bay in southeast Alaska; this one glacial system lost 3030 km^3 of ice, mostly during the nineteenth century [37.163]. Ice loss across the entire region in the twentieth century remained rapid, and *Berthier et al.* [37.162] estimated



that Alaskan glaciers lost $1800 \pm 360 \text{ km}^3$ of ice over 1962–2006, based on differencing of digital elevation models. That rate was probably sustained over the first

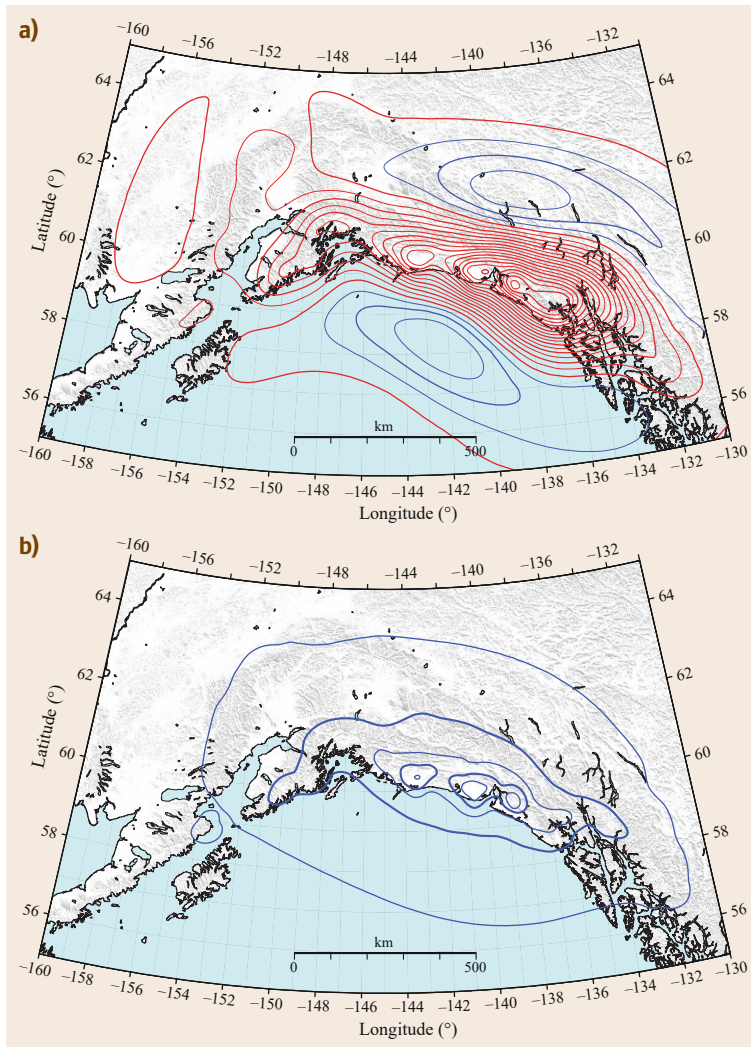


Fig. 37.25a,b Alaska GIA model uplift rates and geoid rates (after [37.58]). **(a)** Uplift and subsidence rates. Contour interval is 2 mm/yr, with *red contours* indicating uplifting regions and *blue contours* representing subsiding regions. **(b)** Geoid change rates, neglecting the redistribution of seawater. Contour intervals are 2 mm/yr, with the same color scheme as the vertical motion rates

half of the twentieth century as well, and accelerated over the last two decades [37.163].

These large ice losses cause extremely rapid uplift rates, which have been observed by repeated GPS measurement since the late 1990s. Because the ice loading model and the Earth response are both known, these data can be used to estimate the viscosity structure of the Earth. While the Laurentide and Fennoscandian ice sheets were emplaced on very old, cold, and rigid continental shields, the Earth structure beneath coastal Alaska is very different. This region was a subduction zone for a long time, until ≈ 30 –50 million years ago, and is characterized by a thin elastic lithosphere (≈ 50 km) and a low viscosity upper mantle, 110 km thick and $3.7 \cdot 10^{18}$ Pa s [37.58], about two orders of magnitude less than the viscosity inferred at that depth from the large continental ice sheets (Fig. 37.24). The

low viscosity is probably due mainly to hydration of the mantle caused by the long period of subduction. The low viscosity means that the mantle flow occurs much more rapidly than for a higher mantle viscosity, producing faster uplift rates for a shorter length of time. Due to the length scale of the regional Alaska load, the present-day response is very sensitive to the relaxation of this shallow, low viscosity mantle layer, and provides little information about the deeper mantle. Larsen et al. [37.163] fit the GPS velocities to such an Earth model, and demonstrated that the model could fit both the total uplift over the last two centuries and the present-day uplift rates.

Elliott et al. [37.58] provided an updated version of this model (computed to spherical harmonic degree and order 2048), and the predicted model uplift rates and geoid change rates are shown in Fig. 37.25. Uplift rates

exceed 30 mm/yr in several places, and geoid change rates exceed 5 mm/yr. These calculations do not take into account the impact of redistribution of seawater due to the changing gravity field, which given the scale of the load has only a small impact on the vertical velocities. However, it may have a larger effect on the geoid change rate, so the values shown in Fig. 37.25b are lower bounds on the magnitude of the geoid change rate.

The Patagonian icefield in South America also experiences rapid rates of ice loss, although with less overall mass change than coastal Alaska [37.167]. Accordingly, uplift rates there are also extremely rapid, and the peak uplift rate observed in Patagonia even exceeds the most rapid rate reported for Alaska [37.169]. The tectonic setting and Earth structure for Patagonia is similar to that of Alaska, and the estimated viscosity structure is comparable as well [37.170, 171].

Networks for the Study of the Present Ice Sheets

Over the last several years under the auspices of the POLENET (polar Earth observing network) project, extensive networks of continuous GNSS sites have been established to measure changes in the polar ice sheets in Antarctica and Greenland (Fig. 37.26). A few stations

had been installed in Greenland and Antarctica by the mid-1990s, but a major expansion began around 2007, at the time of the international polar year. All of these networks represent broad international collaborations.

The Greenland GPS network (GNET) in Greenland (Fig. 37.26a) was installed mostly in 2007–2008, and has documented rapid uplift and variations in uplift rate that correlates with large-scale melting events on the Greenland ice sheet [37.147, 172]. The longest-running sites in Greenland also document clear evidence for a significant change in uplift rates in the early 2000s, which is related to a significant increase in overall mass wastage from the Greenland ice sheet [37.147, 173].

Rapid uplift rates from the Antarctic peninsula were reported by Dietrich et al. [37.174], although they were unsure of the significance of that result at the time, mostly because of concern about large vertical velocity uncertainties. Thomas et al. [37.175] analyzed GPS data from more than 50 stations across Antarctica, and estimated a more complete and precise vertical velocity field. On the Antarctic Peninsula, they found evidence for a significant increase in uplift rates after 2002–2003, which corresponded to the time of the breakup of the Larsen B ice shelf. The breakup of the ice shelf triggered acceleration and thinning of the tributary glaciers that feed into it [37.176, 177], which in turn resulted in

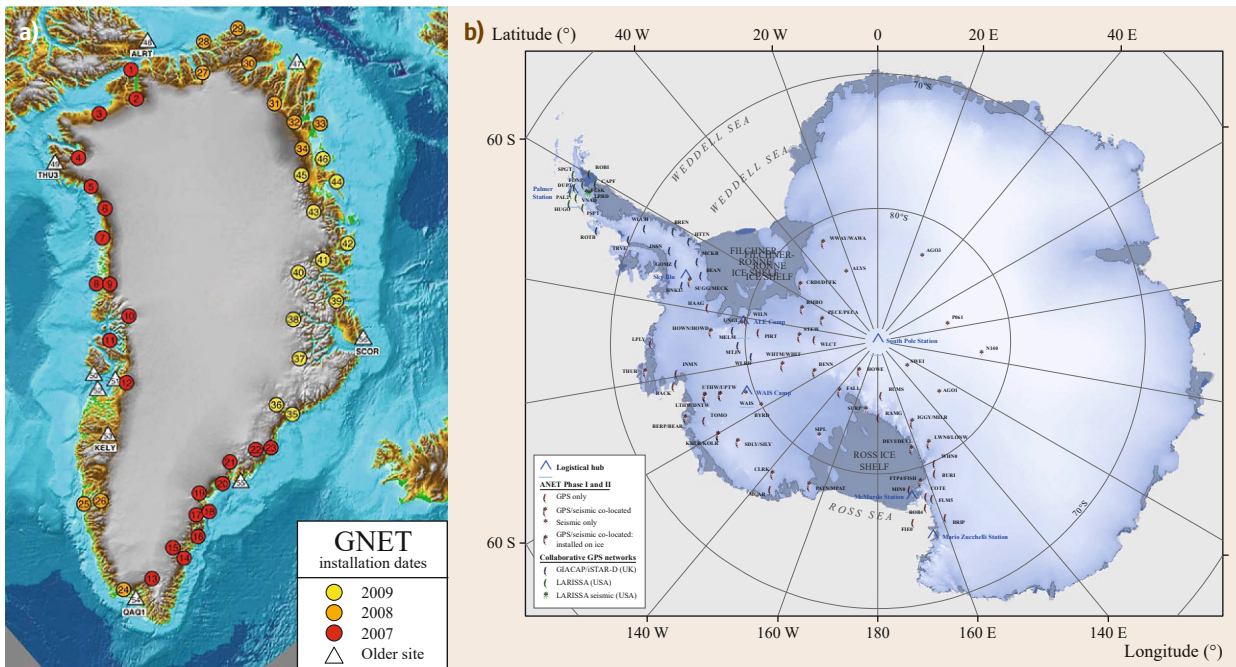


Fig. 37.26a,b Maps of recently installed continuous GNSS networks to study changes in the polar ice sheets. (a) GNET in Greenland, with sites color coded based on their installation date. (b) A-NET in Antarctica, with symbols indicating subnetworks from different contributing organizations. A-NET includes seismic installations as well (after [37.168], courtesy of the POLENET project)

more rapid uplift [37.178]. *Nield et al.* [37.178] added data from six additional continuous GNSS stations to the data set from [37.175] and developed a GIA model to explain the observed uplift rates. They found that the upper mantle viscosity was in the range of $6 \cdot 10^{17}$ – $2 \cdot 10^{18}$ Pa s, making the Antarctic peninsula another example of a low viscosity region.

Several GIA and viscosity models exist for the main body of Antarctica, and some of the viscosity models are shown in Fig. 37.24 [37.165, 179–181]. These models differ considerably (Fig. 37.24), because constraints on the loading history are quite limited, which means that both the viscosity structure and loading his-

tory must be determined mainly from the present-day uplift rates. This leaves the problem poorly constrained, with differences in load changes since LGM differs by as much as 70%. Models that have higher viscosities also have larger ice load changes, to produce comparable present-day deformation. However, in general we can conclude that the main body of Antarctica has a viscosity structure similar to the mid-continental structure found for the Laurentide and Fennoscandian ice sheets. Uncertainties in the magnitude of the GIA signal due to post-LGM deglaciation are of critical importance in the correction of GRACE gravity change data to measure present-day ice mass balance.

37.8 The Multi-GNSS Future

The examples shown in this chapter have all come from a single GNSS system, GPS. Thus far, with the demanding levels of precision and accuracy required for geodynamic applications, GPS+GLONASS solutions have not been demonstrated to be superior to GPS-only solutions. However, this may change with improved Russian Global Navigation Satellite System (GLONASS) modeling, and with the upcoming GLONASS-K2 constellation, which will broadcast code division multiple access (CDMA) signals, more like GPS. Soon, other GNSS systems such as Galileo and Beidou will be equally mature. Details of the various GNSS constellations are given in Chaps. 7–11 of this handbook.

In the near future, multiple full satellite constellations will be available along with a global multi-GNSS tracking network and multi-GNSS orbit and clock products. What impacts can be expected on geodynamic studies? Three areas of impact seem likely:

1. Improvement of precision and accuracy through the addition of independent data.
2. Identification and reduction of systematic errors in GPS through comparison of multi-GNSS solutions.
3. A significant enhancement of kinematic and subdaily positioning capabilities resulting from the improved satellite geometry (more satellites in the sky, and in different parts of the sky).

The first benefit of additional mature multi-GNSS constellations will be the improvement of precision and accuracy through the addition of independent data. A solution using N GNSS constellations with equal precision and accuracy should have its uncertainties reduced by a factor of $N^{1/2}$ relative to a solution using a single constellation. Improvements in combined multi-GNSS solutions could be larger than that, because estimates of parameters like the troposphere delay that

are common to all constellations will be improved, and the greater number of simultaneous satellites in view will also help separate the tropospheric delay and vertical coordinate parameters.

It is likely that systematic errors due to mismodeling of various effects will impact solutions differently for different GNSS constellations. Thus, a careful comparison of independent GNSS solutions or a multi-GNSS combination solution will aid in the identification and reduction of these errors. For example, progress has been made in identifying the cause of the draconitic periodic errors in GPS solutions that were discussed in Sect. 37.1.3. However, even the most recent study of [37.18] did not eliminate these errors, which means that further modeling improvements are needed. Multi-GNSS solutions may reveal incompatibilities and errors in models that can then be addressed. Site-specific errors such as multipath may be different for different constellations, potentially allowing enhanced calibration and removal of such errors.

The improvement in kinematic and subdaily solutions using multiple GNSS constellations may be greater than the improvement in daily solutions. The primary reason for this is the enhanced observing geometry that results from having more satellites in view simultaneously. In general, multi-GNSS observations will provide data from satellites in different parts of the sky, so that multi-GNSS observations will have more uniform sky coverage at all times than a single constellation can provide. In addition to improved precision, the greater geometric strength and more uniform sky coverage will allow for enhanced estimation of tropospheric delays, which often must be modeled in a fairly simple fashion in kinematic solutions. Although the ultimate accuracy of future multi-GNSS kinematic positioning is not easy to predict, it is likely to be signif-

icantly better than kinematic positioning with a single GNSS constellation.

Considering all of the above factors, multi-GNSS solutions eventually should reduce the uncertainty of a daily GNSS position estimate to well below the 1 mm level for the horizontal, and to close to 1 mm for the vertical. This will aid in the detection and modeling of very small motions, and in distinguishing between models that make similar predictions. The improved precision will aid in the study of loading deformation and vertical motion relevant to sea level rise, which both involve small signals. It is likely that the improved precision and accuracy of future multi-GNSS systems

will improve our ability to measure deformation processes and distinguish between competing geodynamic models, and it is possible that previously unrecognized motions and deformations of the Earth will be identified.

Acknowledgments. The author thanks Kimberly de Grandpre and Shanshan Li for providing comments on an early draft from a student's perspective. Many figures in this chapter were created using the Generic Mapping Tools (GMT) developed and maintained by Paul Wessel, Walter H. F. Smith, Remko Scharroo, Joaquim Luis, and Florian Wobbe.

References

- 37.1 M. Bonafede, J. Strehlau, A.R. Ritsema: Geophysical and structural aspects of fault mechanics – A brief historical review, *Terra Nova* **4**(4), 458–463 (1992)
- 37.2 H.F. Reid: Permanent displacements of the ground. In: *The California Earthquake of April 18, 1906*, Vol. II, ed. by S.E.I. Commission (Carnegie Inst. Wash., Washington DC 1910) pp. 16–28
- 37.3 H.F. Reid: The elastic-rebound theory of earthquakes, *Bull. Dept. Geol.* **6**(9), 413–444 (1911)
- 37.4 G.K. Gilbert: A theory of the earthquakes of the Great basin, with a practical application, *Am. J. Sci.* **27**(157), 49–53 (1884)
- 37.5 Z. Altamimi, X. Collilieux, J. Legrand, B. Garayt, C. Boucher: ITRF2005: A new release of the international terrestrial reference frame based on time series of station positions and earth orientation parameters, *J. Geophys. Res.* **112**(B004949), 1–19 (2007)
- 37.6 P. Segall: *Earthquake and Volcano Deformation* (Princeton Univ. Press, Princeton 2010)
- 37.7 E.M. Hill, J.L. Davis, P. Elósegui, B.P. Wernicke, E. Malikowski, N.A. Niemi: Characterization of site-specific GPS errors using a short-baseline network of braced monuments at Yucca mountain, southern Nevada, *J. Geophys. Res. Solid Earth* **114**(B11402), 1–13 (2009)
- 37.8 R.A. Bennett, S. Hreinsdóttir, M.S. Velasco, N.P. Fay: GPS constraints on vertical crustal motion in the northern basin and range, *Geophys. Res. Lett.* **34**(L22319), 1–5 (2007)
- 37.9 M.S. Bos, R.M.S. Fernandes, S.D.P. Williams, L. Bastos: Fast error analysis of continuous GNSS observations with missing data, *J. Geod.* **87**(4), 351–360 (2013)
- 37.10 M. Hackl, R. Malservisi, U. Hugentobler, R. Wonacott: Estimation of velocity uncertainties from GPS time series: Examples from the analysis of the South African TrigNet network, *J. Geophys. Res. Solid Earth* **116**(B11404), 1–12 (2011)
- 37.11 A. Santamaría-Gómez, M.-N. Bouin, X. Collilieux, G. Wöppelmann: Correlated errors in GPS position time series: Implications for velocity estimates, *J. Geophys. Res. Solid Earth* **116**(B01405), 1–14 (2011)
- 37.12 D. Dong, T. Yunck, M. Heflin: Origin of the international terrestrial reference frame, *J. Geophys. Res. Solid Earth* **108**(B4), ETG 8.1–8.10 (2003), doi:10.1029/2002JB002035
- 37.13 D.F. Argus, R.G. Gordon, M.B. Heflin, C. Ma, R.J. Eanes, P. Willis, W.R. Peltier, S.E. Owen: The angular velocities of the plates and the velocity of Earth's centre from space geodesy, *Geophys. J. Int.* **180**(3), 913–960 (2010)
- 37.14 D.F. Argus: Uncertainty in the velocity between the mass center and surface of Earth, *J. Geophys. Res. Solid Earth* **117**(B10), 1–15 (2012)
- 37.15 X. Wu, X. Collilieux, Z. Altamimi, B.L.A. Vermeersen, R.S. Gross, I. Fukumori: Accuracy of the international terrestrial reference frame origin and earth expansion, *Geophys. Res. Lett.* **38**(L13304), 1–5 (2011)
- 37.16 J. Ray, Z. Altamimi, X. Collilieux, T. van Dam: Anomalous harmonics in the spectra of GPS position estimates, *GPS Solutions* **12**(1), 55–64 (2008)
- 37.17 X. Collilieux, L. Métivier, Z. Altamimi, T. van Dam, J. Ray: Quality assessment of GPS reprocessed terrestrial reference frame, *GPS Solutions* **15**(3), 219–231 (2011)
- 37.18 C.J. Rodríguez-Solano, U. Hugentobler, P. Steigenberger, M. Bloßfeld, M. Fritsche: Reducing the draconitic errors in GNSS geodetic products, *J. Geod.* **88**(6), 559–574 (2014)
- 37.19 J. Griffiths, J.R. Ray: Sub-daily alias and draconitic errors in the IGS orbits, *GPS Solutions* **17**(3), 413–422 (2013)
- 37.20 M.A. King, C.S. Watson: Long GPS coordinate time series: Multipath and geometry effects, *J. Geophys. Res. Solid Earth* **115**(B04403), 1–23 (2010)
- 37.21 A.R. Amiri-Simkooei: On the nature of GPS draconitic year periodic pattern in multivariate position time series, *J. Geophys. Res. Solid Earth* **118**(15), 2500–2511 (2013)

- 37.22 R. Zou, J.T. Freymueller, K. Ding, S. Yang, Q. Wang: Evaluating seasonal loading models and their impact on global and regional reference frame alignment, *J. Geophys. Res. Solid Earth* **119**(2), 1337–1358 (2014)
- 37.23 B.A.C. Ambrosius, G. Beutler, G. Blewitt, R.E. Neilan: The role of GPS in the WEGENER project, *J. Geodyn.* **25**(3), 213–240 (1998)
- 37.24 C. Bruyninx: The EUREF permanent network: A multi-disciplinary network serving surveyors as well as scientists, *Geoinformatics* **7**(5), 32–35 (2004)
- 37.25 M. Heflin, W. Bertiger, G. Blewitt, A. Freedman, K. Hurst, S. Lichten, U. Lindqwister, Y. Vigue, F. Webb, T. Yunck, J. Zumberge: Global geodesy using GPS without fiducial sites, *Geophys. Res. Lett.* **19**(2), 131–134 (1992)
- 37.26 G. Beutler, I.I. Mueller, R.E. Neilan: The international GPS service for geodynamics (IGS): The story. In: *GPS Trends in Precise Terrestrial, Airborne, and Spaceborne Applications*, ed. by G. Beutler, G. Hein, W.G. Melbourne, G. Seeber (Springer, Berlin 1996) pp. 3–13
- 37.27 S. Hreinsdóttir, J.T. Freymueller, R. Bürgmann, J. Mitchell: Coseismic deformation of the 2002 Denali fault earthquake: Insights from GPS measurements, *J. Geophys. Res. Solid Earth* **111**(B03308), 1–18 (2006)
- 37.28 Q. Wang, X. Qiao, Q. Lan, J. Freymueller, S. Yang, C. Xu, Y. Yang, X. You, K. Tan, G. Chen: Rupture of deep faults in the 2008 Wenchuan earthquake and uplift of the Longmen Shan, *Nat. Geosci.* **4**(9), 634–640 (2011)
- 37.29 H. Tsuji, M.O. Murakami: Japanese regional GPS tracking network for geodesy and geodynamics. In: *Permanent Satellite Tracking Networks for Geodesy and Geodynamics*, ed. by G.L. Mader (Springer, Vienna 1993) pp. 161–166
- 37.30 J.M. Johansson, J.L. Davis, H.-G. Scherneck, G.A. Milne, M. Vermeer, J.X. Mitrovica, B. Bennett, R.A. Jonsson, G. Elgered, P. Elósegui, H. Koivula, M. Poutanen, B.O. Rönnäng, I.I. Shapiro: Continuous GPS measurements of postglacial adjustment in Fennoscandia: 1. Geodetic results, *J. Geophys. Res. Solid Earth* **107**(B8), 3.1–3.28 (2002)
- 37.31 M. Lidberg, J.M. Johansson, H.-G. Scherneck, G.A. Milne: Recent results based on continuous GPS observations of the GIA process in Fennoscandia from BIFROST, *J. Geodyn.* **50**(1), 8–18 (2010)
- 37.32 Z. Shen, D.D. Jackson, Y. Feng, M. Cline, M. Kim, P. Fang, Y. Bock: Postseismic deformation following the Landers earthquake, California, 28 June 1992, *Bull. Seismol. Soc. Am.* **84**(3), 780–791 (1994)
- 37.33 A.M. Freed, R. Bürgmann: Evidence of power-law flow in the Mojave desert mantle, *Nature* **430**(6999), 548–551 (2004)
- 37.34 K.W. Hudnut, Y. Bock, M. Cline, P. Fang, Y. Feng, J. Freymueller, X. Ge, W.K. Gross, D. Jackson, M. Kim, N.E. King, J. Langbein, S.C. Larsen, M. Lisowski, Z.-K. Shen, J. Svarc, J. Zhang: Co-seismic displacements of the 1992 Landers earthquake sequence, *Bull. Seismol. Soc. Am.* **84**(3), 625–645 (1994)
- 37.35 J. Freymueller, N.E. King, P. Segall: The co-seismic slip distribution of the Landers earthquake, *Bull. Seismol. Soc. Am.* **84**(3), 646–659 (1994)
- 37.36 K.W. Hudnut, Y. Bock, J.E. Galetzka, F.H. Webb, W.H. Young: The southern California integrated GPS network (SCIGN), *Proc. 10th FIG Int. Symp. Deform. Meas.*, Orange (FIG, Copenhagen 2001) pp. 19–22
- 37.37 T. Sagiya: A decade of GEONET: 1994–2003–The continuous GPS observation in Japan and its impact on earthquake studies, *Earth Planets Space* **56**(8), xxix–xlII (2004)
- 37.38 G.F. Sella, T.H. Dixon, A. Mao: REVEL: A model for recent plate velocities from space geodesy, *J. Geophys. Res. Solid Earth* **107**(B4), ETG 11.1–11.31 (2002)
- 37.39 R. McCaffrey, A.I. Qamar, R.W. King, R. Wells, G. Khazaradze, C.A. Williams, C.W. Stevens, J.J. Vollick, P.C. Zwick: Fault locking, block rotation and crustal deformation in the Pacific northwest, *Geophys. J. Int.* **169**(3), 1315–1340 (2007)
- 37.40 J.T. Freymueller: GPS, tectonic geodesy. In: *Encyclopedia of Solid Earth Geophysics*, ed. by H.K. Gupta (Springer, Berlin 2011) pp. 431–449
- 37.41 G.F. Sella, S. Stein, T.H. Dixon, M. Craymer, T.S. James, S. Mazzotti, R.K. Dokka: Observation of glacial isostatic adjustment in stable North America with GPS, *Geophys. Res. Lett.* **34**(L02306), 1–6 (2007)
- 37.42 E. Calais, J.Y. Han, C. DeMets, J.M. Nocquet: Deformation of the North American plate interior from a decade of continuous GPS measurements, *J. Geophys. Res. Solid Earth* **111**(B06402), 1–13 (2006)
- 37.43 K.M. Larson, J.T. Freymueller, S. Philipsen: Global plate velocities from the global positioning system, *J. Geophys. Res. Solid Earth* **102**(B5), 9961–9981 (1997)
- 37.44 E.O. Norabuena, T.H. Dixon, S. Stein, C.G.A. Harrison: Decelerating Nazca–South America and Nazca–Pacific plate motions, *Geophys. Res. Lett.* **26**(22), 3405–3408 (1999)
- 37.45 W. Thatcher: How the continents deform: The evidence from tectonic geodesy, *Annu. Rev. Earth Planet. Sci.* **37**, 237–262 (2009)
- 37.46 W. Gan, P. Zhang, Z.-K. Shen, Z. Niu, M. Wang, Y. Wan, D. Zhou, J. Cheng: Present-day crustal motion within the Tibetan plateau inferred from GPS measurements, *J. Geophys. Res. Solid Earth* **112**(B08416), 1–14 (2007)
- 37.47 C.H. Scholz: *The Mechanics of Earthquakes and Faulting* (Cambridge Univ. Press, Cambridge 2002)
- 37.48 J.C. Savage, R.O. Burford: Accumulation of tectonic strain in California, *Bull. Seismol. Soc. Am.* **60**(6), 1877–1896 (1970)
- 37.49 J.C. Savage: A dislocation model of strain accumulation and release at a subduction zone, *J. Geophys. Res. Solid Earth* **88**(B6), 4984–4996 (1983)
- 37.50 W. Thatcher: Strain accumulation on the northern San Andreas fault zone since 1906, *J. Geophys.*

- Res. **80**(35), 4873–4880 (1975)
- 37.51 W. Thatcher, J.B. Rundle: A viscoelastic coupling model for the cyclic deformation due to periodically repeated Earthquakes at subduction zones, *J. Geophys. Res. Solid Earth* **89**(B9), 7631–7640 (1984)
- 37.52 J.T. Freymueller: Active tectonics of plate boundary zones and the continuity of plate boundary deformation from Asia to North America, *Curr. Sci.* **99**(12), 1719–1732 (2010)
- 37.53 A.J. Haines, W.E. Holt: A procedure to obtain the complete horizontal motions within zones of distributed deformation from the inversion of strain rate data, *J. Geophys. Res.* **98**(B7), 12057–12082 (1993)
- 37.54 T. Candela, F. Renard, Y. Klinger, K. Mair, J. Schmittbuhl, E.E. Brodsky: Roughness of fault surfaces over nine decades of length scales, *J. Geophys. Res. Solid Earth* **117**(B08409), 1–30 (2012)
- 37.55 Y. Okada: Surface deformation due to shear and tensile faults in a half-space, *Bull. Seismol. Soc. Am.* **75**(4), 1135–1154 (1985)
- 37.56 R. McCaffrey: Crustal block rotations and plate coupling. In: *Plate Boundary Zones, AGU Geodynamics Series 30*, ed. by S. Stein, J. Freymueller (AGU, Washington DC 2002) pp. 101–122
- 37.57 B.J. Meade, B.H. Hager: Block models of crustal motion in southern California constrained by GPS measurements, *J. Geophys. Res. Solid Earth* **110**(B03403), 1–19 (2005)
- 37.58 J.L. Elliott, C.F. Larsen, J.T. Freymueller, R.J. Motyka: Tectonic block motion and glacial isostatic adjustment in southeast Alaska and adjacent Canada constrained by GPS measurements, *J. Geophys. Res. Solid Earth* **115**(B09407), 1–21 (2010)
- 37.59 J. Elliott, J.T. Freymueller, C.F. Larsen: Active tectonics of the St. Elias orogen, Alaska, observed with GPS measurements, *J. Geophys. Res. Solid Earth* **118**(10), 5625–5642 (2013)
- 37.60 M.A. Langstaff, B.J. Meade: Edge-driven mechanical microplate models of strike-slip faulting in the Tibetan plateau, *J. Geophys. Res. Solid Earth* **118**(7), 3809–3819 (2013)
- 37.61 J.P. Loveless, B.J. Meade: Geodetic imaging of plate motions, slip rates, and partitioning of deformation in Japan, *J. Geophys. Res. Solid Earth* **115**(B02410), 1–35 (2010)
- 37.62 Q. Chen, J.T. Freymueller, Q. Wang, Z. Yang, C. Xu, J. Liu: A deforming block model for the present-day tectonics of Tibet, *J. Geophys. Res. Solid Earth* **109**(B01403), 1–16 (2004)
- 37.63 E. Parkin: Horizontal crustal movements. In: *The Great Alaska Earthquake of 1964 – Seismology and Geodesy*, ed. by Cot.A. Earthquake (Natl. Acad. Sci., Washington DC 1972) pp. 419–434
- 37.64 S.C. Cohen, J.T. Freymueller: Crustal deformation in the southcentral Alaska subduction zone, *Adv. Geophys.* **47**, 1–63 (2004)
- 37.65 M. Sato, T. Ishikawa, N. Ujihara, S. Yoshida, M. Fujita, M. Mochizuki, A. Asada: Displacement above the hypocenter of the 2011 Tohoku–Oki earthquake, *Science* **332**(6036), 1395 (2011)
- 37.66 T. Lay, C.J. Ammon, H.O. Kanamori, L. Xue, M. Kim: Possible large near-trench slip during the 2011 M (w) 9.0 off the Pacific coast of Tohoku Earthquake, *Earth, Planets Space* **63**(7), 687–692 (2011)
- 37.67 R. Wang, F. Lorenzo-Martín, F. Roth: PSGRN–PSCMP – A new code for calculating co- and post-seismic deformation, geoid and gravity changes based on the viscoelastic-gravitational dislocation theory, *Comput. Geosci.* **32**(4), 527–541 (2006)
- 37.68 F.F. Pollitz: Coseismic deformation from earthquake faulting on a layered spherical Earth, *Geophys. J. Int.* **125**(1), 1–14 (1996)
- 37.69 W. Menke: *Geophysical Data Analysis: Discrete Inverse Theory* (Elsevier Academic, Amsterdam 2012)
- 37.70 R.C. Aster, B. Borchers, C.H. Thurber: *Parameter Estimation and Inverse Problems* (Elsevier Academic, Amsterdam 2012)
- 37.71 P.B. Stark, R.L. Parker: Bounded-variable least-squares: An algorithm and applications, *Comput. Stat.* **10**, 129 (1995)
- 37.72 M.V. Matthews, P. Segall: Estimation of depth-dependent fault slip from measured surface deformation with application to the 1906 San Francisco earthquake, *J. Geophys. Res. Solid Earth* **98**(B7), 12153–12163 (1993)
- 37.73 P. Banerjee, F.F. Pollitz, R. Bürgmann: The size and duration of the Sumatra–Andaman earthquake from far-field static offsets, *Science* **308**(5729), 1769–1772 (2005)
- 37.74 W. Wang, W. Sun, Y. Wu, G. Gu: Modification of fault slip models of the M_w 9.0 Tohoku Earthquake by far field GPS observations, *J. Geodyn.* **75**, 22–33 (2014)
- 37.75 P. Tregoning, R. Burgette, S.C. McClusky, S. Lejune, C.S. Watson, H. McQueen: A decade of horizontal deformation from great earthquakes, *J. Geophys. Res. Solid Earth* **118**(5), 2371–2381 (2013)
- 37.76 R.M. Nikolaidis, Y. Bock, P.J. Jonge, P. Shearer, D.C. Agnew, M. van Domselaar: Seismic wave observations with the global positioning system, *J. Geophys. Res. Solid Earth* **106**(B10), 21897–21916 (2001)
- 37.77 K.M. Larson, P. Bodin, J. Gomberg: Using 1 Hz GPS data to measure deformations caused by the Denali fault earthquake, *Science* **300**(5624), 1421–1424 (2003)
- 37.78 A. Avallone, M. Marzario, A. Cirella, A. Piatanesi, A. Rovelli, C. di Alessandro, E. D’Anastasio, N. D’Agostino, R. Giuliani, M. Mattone: Very high rate (10 Hz) GPS seismology for moderate–magnitude earthquakes: The case of the Mw 6.3 L’Aquila (central Italy) event, *J. Geophys. Res. Solid Earth* **116**(B02305), 1–14 (2011)
- 37.79 Y. Zheng, J. Li, Z. Xie, M.H. Ritzwoller: 5 Hz GPS seismology of the El Mayor–Cucapah earthquake: Estimating the earthquake focal mechanism, *Geophys. J. Int.* **190**(3), 1723–1732 (2012)
- 37.80 H. Yue, T. Lay, J.T. Freymueller, K. Ding, L. Rivera, N.A. Ruppert, K.D. Koper: Supershear rupture of the 5 January 2013 Craig, Alaska (Mw 7.5) earthquake, *J. Geophys. Res. Solid Earth* **118**(11), 5903–

- 5919 (2013)
- 37.81 J.F. Genrich, Y. Bock: Instantaneous geodetic positioning with 10–50 Hz GPS measurements: Noise characteristics and implications for monitoring networks, *J. Geophys. Res. Solid Earth* **111**(B03403), 1–16 (2006)
- 37.82 S. Miyazaki, P. Segall, J. Fukuda, T. Kato: Space time distribution of afterslip following the 2003 Tokachi-oki earthquake: Implications for variations in fault zone frictional properties, *Geophys. Res. Lett.* **31**(L06623), 1–4 (2004)
- 37.83 B.W. Crowell, Y. Bock, D. Melgar: Real-time inversion of GPS data for finite fault modeling and rapid hazard assessment, *Geophys. Res. Lett.* **39**(L09305), 1–6 (2012)
- 37.84 P. Elósegui, J.L. Davis, D. Oberlander, R. Baena, G. Ekström: Accuracy of high-rate GPS for seismology, *Geophys. Res. Lett.* **33**(L11308), 1–4 (2006)
- 37.85 F. Moschas, S. Stiros: PLL bandwidth and noise in 100 Hz GPS measurements, *GPS Solutions* **19**(2), 173–185 (2014)
- 37.86 R. Tu, R. Wang, M. Ge, T.R. Walter, M. Ramatschi, C. Milkereit, D. Bindl, T. Dahm: Cost-effective monitoring of ground motion related to earthquakes, landslides, or volcanic activity by joint use of a single-frequency GPS and a MEMS accelerometer, *Geophys. Res. Lett.* **40**(15), 3825–3829 (2013)
- 37.87 H. Yue, T. Lay, K.D. Koper: En echelon and orthogonal fault ruptures of the 11 April 2012 great intraplate earthquakes, *Nature* **490**(7419), 245–249 (2012)
- 37.88 B.W. Crowell, Y. Bock, M.B. Squibb: Demonstration of earthquake early warning using total displacement waveforms from real-time GPS networks, *Seismol. Res. Lett.* **80**(5), 772–782 (2009)
- 37.89 S.E. Owen, F. Webb, M. Simons, P.A. Rosen, J. Cruz, S. Yun, E.J. Fielding, A.W. Moore, H. Hua, P.S. Agram: The ARIA-EQ project: Advanced rapid imaging and analysis for earthquakes, *Proc. AGU Fall Meet., San Francisco (AGU, Washington DC 2011)* p. 1298
- 37.90 X. Li, G. Dick, M. Ge, S. Heise, J. Wickert, M. Bender: Real-time GPS sensing of atmospheric water vapor: Precise point positioning with orbit, clock, and phase delay corrections, *Geophys. Res. Lett.* **41**(10), 3615–3621 (2014)
- 37.91 X. Li, M. Ge, X. Zhang, Y. Zhang, B. Guo, R. Wang, J. Klotz, J. Wickert: Real-time high-rate coseismic displacement from ambiguity-fixed precise point positioning: Application to earthquake early warning, *Geophys. Res. Lett.* **40**(2), 295–300 (2013)
- 37.92 Y. Ohta, T. Kobayashi, H. Tsushima, S. Miura, R. Hino, T. Takasu, H. Fujimoto, T. Iinuma, K. Tachibana, T. Demachi, T. Sato, M. Ohzono, N. Um: Quasi real-time fault model estimation for near-field tsunami forecasting based on RTK-GPS analysis: Application to the 2011 Tohoku-Oki earthquake (Mw 9.0), *J. Geophys. Res. Solid Earth* **117**(B02311), 1–16 (2012)
- 37.93 D. Melgar, Y. Bock, B.W. Crowell: Real-time centroid moment tensor determination for large earthquakes from local and regional displacement records, *Geophys. J. Int.* **188**(2), 703–718 (2012)
- 37.94 S.E. Minson, J.R. Murray, J.O. Langbein, J.S. Gombert: Real-time inversions for finite fault slip models and rupture geometry based on high-rate GPS data, *J. Geophys. Res. Solid Earth* **119**(4), 3201–3231 (2014)
- 37.95 R. Fang, C. Shi, W. Song, G. Wang, J. Liu: Determination of earthquake magnitude using GPS displacement waveforms from real-time precise point positioning, *Geophys. J. Int.* **196**(1), 461–472 (2014)
- 37.96 B. Gutenberg: Amplitudes of surface waves and magnitudes of shallow earthquakes, *Bull. Seismol. Soc. Am.* **35**(1), 3–12 (1945)
- 37.97 B.W. Crowell, D. Melgar, Y. Bock, J.S. Haase, J. Geng: Earthquake magnitude scaling using seismogeodetic data, *Geophys. Res. Lett.* **40**(23), 6089–6094 (2013)
- 37.98 G. Rogers, H. Dragert: Episodic tremor and slip on the Cascadia subduction zone: The chatter of silent slip, *Science* **300**(5627), 1942–1943 (2003)
- 37.99 S.Y. Schwartz, J.M. Rokosky: Slow slip events and seismic tremor at circum-Pacific subduction zones, *Rev. Geophys.* **45**(RG3004), 1–32 (2007)
- 37.100 S. Ide, G.C. Beroza, D.R. Shelly, T. Uchide: A scaling law for slow earthquakes, *Nature* **447**(7140), 76–79 (2007)
- 37.101 R. Bürgmann, P. Segall, M. Lisowski, J. Svarc: Post-seismic strain following the 1989 Loma Prieta earthquake from GPS and leveling measurements, *J. Geophys. Res. Solid Earth* **102**(B3), 4933–4955 (1997)
- 37.102 J.J. Lienkaemper, J.S. Galehouse, R.W. Simpson: Creep response of the Hayward fault to stress changes caused by the Loma Prieta earthquake, *Science* **276**(5321), 2014–2016 (1997)
- 37.103 R. Bürgmann, G. Dresen: Rheology of the lower crust and upper mantle: Evidence from rock mechanics, geodesy, and field observations, *Annu. Rev. Earth Planet. Sci.* **36**(1), 531–567 (2008)
- 37.104 M. Moreno, D. Melnick, M. Rosenau, J. Baez, J. Klotz, O. Oncken, A. Tassara, J. Chen, K. Bataille, M. Bevis, A. Socquet, J. Bolte, C. Vigny, B. Brooks, I. Ryder, V. Grund, B. Smalley, D. Carrizo, M. Bartsch, H. Hase: Toward understanding tectonic control on the M_w 8.8 2010 Maule Chile earthquake, *Earth Planet. Sci. Lett.* **321**, 152–165 (2012)
- 37.105 Y.N. Lin, A. Sladen, F. Ortega-Culaciati, M. Simons, J.-P. Avouac, E.J. Fielding, B.A. Brooks, M. Bevis, J. Genrich, A. Rietbrock, C. Vigny, R. Smalley, A. Scocquet: Coseismic and postseismic slip associated with the 2010 Maule earthquake, Chile: Characterizing the Arauco peninsula barrier effect, *J. Geophys. Res. Solid Earth* **118**(6), 3142–3159 (2013)
- 37.106 C. Marone, C.B. Raleigh, C.H. Scholz: Frictional behavior and constitutive modeling of simulated

- fault gouge, *J. Geophys. Res. Solid Earth* **95**(B5), 7007–7025 (1990)
- 37.107 Y.-J. Hsu, M. Simons, J.-P. Avouac, J. Galetzka, K. Sieh, M. Chlieh, D. Natawidjaja, L. Prawirodirdjo, Y. Bock: Frictional afterslip following the 2005 Nias–Simeulue earthquake, Sumatra, *Science* **312**(5782), 1921–1926 (2006)
- 37.108 I.A. Johanson, E.J. Fielding, F. Rolandone, R. Bürgmann: Coseismic and postseismic slip of the 2004 Parkfield earthquake from space-geodetic data, *Bull. Seismol. Soc. Am.* **96**(4B), S269–S282 (2006)
- 37.109 C. Kreemer, G. Blewitt, F. Maerten: Co- and postseismic deformation of the 28 March 2005 Nias Mw 8.7 earthquake from continuous GPS data, *Geophys. Res. Lett.* **33**(L07307), 1–4 (2006)
- 37.110 S. Ozawa, T. Nishimura, H. Munekane, H. Suito, T. Kobayashi, M. Tobita, T. Imakiire: Preceding, coseismic and postseismic slips of the 2011 Tohoku earthquake, Japan, *J. Geophys. Res. Solid Earth* **117**(B07404), 1–20 (2012)
- 37.111 K.M. Johnson, J. Fukuda, P. Segall: Challenging the rate–state asperity model: Afterslip following the 2011 M9 Tohoku–oki, Japan, earthquake, *Geophys. Res. Lett.* **39**(L20302), 1–5 (2012)
- 37.112 K. Wang, Y. Hu, J. He: Deformation cycles of subduction earthquakes in a viscoelastic Earth, *Nature* **484**(7394), 327–332 (2012)
- 37.113 M. Bevis, A. Brown: Trajectory models and reference frames for crustal motion geodesy, *J. Geod.* **88**(3), 283–311 (2014)
- 37.114 A.M. Freed, R. Bürgmann, E. Calais, J. Freymueller, S. Hreinsdóttir: Implications of deformation following the 2002 Denali, Alaska, earthquake for postseismic relaxation processes and lithospheric rheology, *J. Geophys. Res. Solid Earth* **111**(B01401), 1–23 (2006)
- 37.115 K. Mogi: Relations between the eruptions of various volcanoes and the deformations of the ground surfaces around them, *Bull. Earthq. Res. Inst. Univ. Tokyo* **36**, 99–134 (1958)
- 37.116 R. Murakami, S. Ozawa, T. Nishimura, T. Tada: A model of magma movements associated with the 2000 eruption of Usu volcano inferred by crustal deformation detected by continuous GPS and other geodetic measurements, *J. Geospatial Inf. Auth.* **95**, 99–105 (2001)
- 37.117 P. Jousset, H. Mori, H. Okada: Elastic models for the magma intrusion associated with the 2000 eruption of Usu Volcano, Hokkaido, Japan, *J. Volcanol. Geotherm. Res.* **125**(1), 81–106 (2003)
- 37.118 J.J. Dvorak, D. Dzurlis: Variations in magma supply rate at Kilauea volcano, Hawaii, *J. Geophys. Res. Solid Earth* **98**(B12), 22255–22268 (1993)
- 37.119 J.J. Dvorak, D. Dzurlis: Volcano geodesy: The search for magma reservoirs and the formation of eruptive vents, *Rev. Geophys.* **35**(3), 343–384 (1997)
- 37.120 T. Fournier, J. Freymueller, P. Cervelli: Tracking magma volume recovery at Okmok volcano using GPS and an unscented Kalman filter, *J. Geophys. Res. Solid Earth* **114**(B02405), 1–18 (2009)
- 37.121 J.F. Larsen, C.J. Nye, M.L. Coombs, M. Tilman, P. Izbekov, C. Cameron: Petrology and geochemistry of the 2006 eruption of Augustine Volcano. In: *The 2006 Eruption of Augustine Volcano, Alaska. US Geological Survey, Professional Paper 1769*, ed. by J.A. Power, M.L. Coombs, J.T. Freymueller (US Geological Survey, Washington DC 2006) pp. 335–382
- 37.122 A. Burgisser, G.W. Bergantz: A rapid mechanism to remobilize and homogenize highly crystalline magma bodies, *Nature* **471**(7337), 212–215 (2011)
- 37.123 K.M. Larson, M. Poland, A. Miklius: Volcano monitoring using GPS: Developing data analysis strategies based on the June 2007 Kilauea Volcano intrusion and eruption, *J. Geophys. Res. Solid Earth* **115**(B07406), 1–10 (2010)
- 37.124 K.M. Larson: A new way to detect volcanic plumes, *Geophys. Res. Lett.* **40**(11), 2657–2660 (2013)
- 37.125 R. Grapenthin, J.T. Freymueller, A.M. Kaufman: Geodetic observations during the 2009 eruption of Redoubt Volcano, Alaska, *J. Volcanol. Geotherm. Res.* **259**, 115–132 (2013)
- 37.126 N.T. Penna, M.P. Stewart: Aliased tidal signatures in continuous GPS height time series, *Geophys. Res. Lett.* **30**(23), SDE 1.1–1.4 (2003)
- 37.127 N.T. Penna, M.A. King, M.P. Stewart: GPS height time series: Short-period origins of spurious long-period signals, *J. Geophys. Res. Solid Earth* **112**(B02402), 1–19 (2007)
- 37.128 J. Kusche, E.J.O. Schrama: Surface mass redistribution inversion from global GPS deformation and gravity recovery and climate experiment (GRACE) gravity data, *J. Geophys. Res. Solid Earth* **110**(B09409), 1–14 (2005)
- 37.129 R.J. Blakely: *Potential Theory in Gravity and Magnetic Applications* (Cambridge Univ. Press, Cambridge 1996)
- 37.130 T. van Dam, J. Wahr, D. Lavallée: A comparison of annual vertical crustal displacements from GPS and gravity recovery and climate experiment (GRACE) over Europe, *J. Geophys. Res. Solid Earth* **112**(B03404), 1–11 (2007)
- 37.131 W.E. Farrell: Deformation of the Earth by surface loads, *Rev. Geophys.* **10**(3), 761–797 (1972)
- 37.132 Y. Fu, D.F. Argus, J.T. Freymueller, M.B. Heflin: Horizontal motion in elastic response to seasonal loading of rain water in the Amazon basin and monsoon water in southeast Asia observed by GPS and inferred from GRACE, *Geophys. Res. Lett.* **40**(23), 6048–6053 (2013)
- 37.133 G. Spada: *The Theory Behind TABOO* (Samizdat, White River Junction 2003)
- 37.134 G. Spada, V.R. Barletta, V. Klemann, R.E.M. Riva, Z. Martinec, P. Gasperini, B. Lund, D. Wolf, L.L.A. Vermeersen, M.A. King: A benchmark study for glacial isostatic adjustment codes, *Geophys. J. Int.* **185**(1), 106–132 (2011)
- 37.135 X. Collilieux, Z. Altamimi, D. Coulot, T. van Dam, J. Ray: Impact of loading effects on determination of the international terrestrial reference frame, *Adv. Space Res.* **45**(1), 144–154 (2010)

- 37.136 X. Collilieux, T. van Dam, J. Ray, D. Coulot, L. Métivier, Z. Altamimi: Strategies to mitigate aliasing of loading signals while estimating GPS frame parameters, *J. Geod.* **86**(1), 1–14 (2012)
- 37.137 K. Heki: Seasonal modulation of interseismic strain buildup in northeastern Japan driven by snow loads, *Science* **293**(5527), 89–92 (2001)
- 37.138 K. Heki: Snow load and seasonal variation of earthquake occurrence in Japan, *Earth Planet. Sci. Lett.* **207**(1), 159–164 (2003)
- 37.139 D. Dong, P. Fang, Y. Bock, M.K. Cheng, S. Miyazaki: Anatomy of apparent seasonal variations from GPS-derived site position time series, *J. Geophys. Res. Solid Earth* **107**(B4), ETG 9–1–ETG 9–16 (2002), doi:[10.1029/2001JB000573](https://doi.org/10.1029/2001JB000573)
- 37.140 G. Blewitt, D. Lavallée, P. Clarke, K. Nurutdinov: A new global mode of Earth deformation: Seasonal cycle detected, *Science* **294**(5550), 2342–2345 (2001)
- 37.141 P. Tregoning, C. Watson, G. Ramillien, H. McQueen, J. Zhang: Detecting hydrologic deformation using GRACE and GPS, *Geophys. Res. Lett.* **36**(L15401), 1–6 (2009)
- 37.142 S. Nahmani, O. Bock, M. Bouin, A. Santamaría-Gómez, J.-P. Boy, X. Collilieux, L. Métivier, I. Panet, P. Genthon, C. Linage, G. Woepellmann: Hydrological deformation induced by the west African monsoon: Comparison of GPS, GRACE and loading models, *J. Geophys. Res. Solid Earth* **117**(B05409), 1–16 (2012)
- 37.143 Y. Fu, J.T. Freymueller: Seasonal and long-term vertical deformation in the Nepal Himalaya constrained by GPS and GRACE measurements, *J. Geophys. Res. Solid Earth* **117**(B03407), 1–14 (2012)
- 37.144 Y. Fu, J.T. Freymueller, T. Jensen: Seasonal hydrological loading in southern Alaska observed by GPS and GRACE, *Geophys. Res. Lett.* **39**(L15310), 1–5 (2012)
- 37.145 J.L. Davis, P. Elósegui, J.X. Mitrovica, M.E. Tamisiea: Climate-driven deformation of the solid Earth from GRACE and GPS, *Geophys. Res. Lett.* **31**(L24605), 1–4 (2004)
- 37.146 M.S. Steckler, S.L. Nooner, S.H. Akhter, S.K. Chowdhury, S. Bettadpur, L. Seeber, M.G. Kogan: Modeling Earth deformation from monsoonal flooding in Bangladesh using hydrographic, GPS, and gravity recovery and climate experiment (GRACE) data, *J. Geophys. Res. Solid Earth* **115**(B08407), 1–18 (2010)
- 37.147 M. Bevis, J. Wahr, S.A. Khan, F.B. Madsen, A. Brown, M. Willis, E. Kendrick, P. Knudsen, J.E. Box, T. van Dam, D.J. Caccamise II, B. Johns, T. Nylen, R. Abbott, S. White, J. Miner, R. Forsberg, H. Zhou, J. Wang, T. Wilson, D. Bromwich, O. Francis: Bedrock displacements in Greenland manifest ice mass variations, climate cycles and climate change, *Proc. Natl. Acad. Sci.* **109**(30), 11944–11948 (2012)
- 37.148 T. van Dam: 3-dimensional surface displacements derived from the GRACE dealiasing products (2012) <http://geophy.uni.lu/ggfc-combination.html>
- 37.149 D. Stammer, C. Wunsch, R. Giering, C. Eckert, P. Heimbach, J. Marotzke, A. Adcroft, C.N. Hill, J. Marshall: Global ocean circulation during 1992–1997, estimated from ocean observations and a general circulation model, *J. Geophys. Res. Oceans* **107**(C9), 1.1–1.27 (2002)
- 37.150 T.M. van Dam, J.M. Wahr: Displacements of the Earth's surface due to atmospheric loading: Effects on gravity and baseline measurements, *J. Geophys. Res. Solid Earth* **92**(B2), 1281–1286 (1987)
- 37.151 M. Rodell, P.R. Houser, U. Jambor, J. Gottschalk, K. Mitchell, C.J. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, J.K. Entin, J.P. Walker, D. Lohmann, D. Toll: The global land data assimilation system, *Bull. Am. Meteorol. Soc.* **85**(3), 381–394 (2004)
- 37.152 M.M. Rienecker, M.J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M.G. Bosilovich, S.D. Schubert, L. Takacs, G.K. Kim, S. Bloom, J. Chen, D. Collins, A. Conaty, A. da Silva, W. Gu, J. Joiner, R.D. Koster, R. Lucchesi, A. Molod, T. Owens, S. Pawson, P. Pegion, C.R. Redder, R. Reichle, F.R. Robertson, A.G. Ruddick, M. Sienkiewicz, J. Wollen: MERRA: NASA's modern-era retrospective analysis for research and applications, *J. Clim.* **24**(14), 3624–3648 (2011)
- 37.153 Z. Li, T. van Dam, X. Collilieux, Z. Altamimi, J. Ray, P. Rebischung, S. Nahmani: Quality evaluation of the weekly vertical loading effects induced from continental water storage models. In: *IAG 150 Years. Vol. 143 of International Association of Geodesy Symposia*, ed. by C. Rizos, P. Willis (Springer, Heidelberg 2016) pp. 673–679
- 37.154 W.R. Peltier: Global glacial isostasy and the surface of the ice-age Earth: The ICE-5G (VM2) model and GRACE, *Annu. Rev. Earth Planet. Sci.* **32**, 111–149 (2004)
- 37.155 M.A. Toscano, W.R. Peltier, R. Drummond: ICE-5G and ICE-6G models of postglacial relative sea-level history applied to the Holocene coral reef record of northeastern St. Croix, US V.I.: Investigating the influence of rotational feedback on GIA processes at tropical latitudes, *Quaternary Sci. Rev.* **30**(21), 3032–3042 (2011)
- 37.156 G.A. Milne, J.X. Mitrovica, H.-G. Scherneck, J.L. Davis, J.M. Johansson, H. Koivula, M. Vermeer: Continuous GPS measurements of postglacial adjustment in Fennoscandia: 2. Modeling results, *J. Geophys. Res. Solid Earth* **109**(B2), ETG 3.1–3.28 (2004)
- 37.157 M. Lidberg, J.M. Johansson, H.-G. Scherneck, J.L. Davis: An improved and extended GPS-derived 3-D velocity field of the glacial isostatic adjustment (GIA) in Fennoscandia, *J. Geod.* **81**(3), 213–230 (2007)
- 37.158 A.M. Tushingham, W.R. Peltier: Ice-3G: A new global model of Late Pleistocene deglaciation based upon geophysical predictions of postglacial relative sea level change, *J. Geophys. Res. Solid Earth* **96**(B3), 4497–4523 (1991)

- 37.159 M.B. Dyurgerov, M.F. Meier: *Glaciers and the Changing Earth System: A 2004 Snapshot* (Institute of Arctic and Alpine Research, Univ. Colorado, Boulder 2005)
- 37.160 C.F. Larsen, R.J. Motyka, A.A. Arendt, K.A. Echelmeyer, P.E. Geissler: Glacier changes in southeast Alaska and northwest British Columbia and contribution to sea level rise, *J. Geophys. Res. Earth Surface* **112**(F01007), 1–11 (2007)
- 37.161 P. Lemke, J. Ren, R.B. Alley, I. Allison, J. Carrasco, G. Flato, Y. Fujii, G. Kaser, P.W. Mote, R.H. Thomas, T. Zhang: Observations: Changes in snow, ice and frozen ground. In: *Climate Change 2007: The Physical Science Basis*, ed. by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, H.L. Miller (Cambridge Univ. Press, Cambridge 2007) pp. 337–383
- 37.162 E. Berthier, E. Schiefer, G.K.C. Clarke, B. Menounos, F. Rémy: Contribution of Alaskan glaciers to sea-level rise derived from satellite imagery, *Nat. Geosci.* **3**(2), 92–95 (2010)
- 37.163 C.F. Larsen, R.J. Motyka, J.T. Freymueller, K.A. Echelmeyer, E.R. Ivins: Rapid viscoelastic uplift in southeast Alaska caused by post-Little Ice Age glacial retreat, *Earth Planet. Sci. Lett.* **237**(3), 548–560 (2005)
- 37.164 W.R. Peltier, R. Drummond: Rheological stratification of the lithosphere: A direct inference based upon the geodetically observed pattern of the glacial isostatic adjustment of the North American continent, *Geophys. Res. Lett.* **35**(L16314), 1–5 (2008)
- 37.165 D.F. Argus, W.R. Peltier, R. Drummond, A.W. Moore: The Antarctica component of postglacial rebound model ICE-6G_C (VM5a) based on GPS positioning, exposure age dating of ice thicknesses, and relative sea level histories, *Geophys. J. Int.* **198**(1), 537–563 (2014)
- 37.166 K. Latychev, J.X. Mitrovica, M.E. Tamisiea, J. Tromp, R. Moucha: Influence of lithospheric thickness variations on 3-D crustal velocities due to glacial isostatic adjustment, *Geophys. Res. Lett.* **32**(L01304), 1–4 (2005)
- 37.167 M.J. Willis, A.K. Melkonian, M.E. Pritchard, A. Rivera: Ice loss from the southern Patagonian ice field, South America, between 2000 and 2012, *Geophys. Res. Lett.* **39**(L17501), 1–6 (2012)
- 37.168 POLENET – The Polar Earth Observation Network, <http://polenet.org/>
- 37.169 R. Dietrich, E.R. Ivins, G. Casassa, H. Lange, J. Wendt, M. Fritsche: Rapid crustal uplift in Patagonia due to enhanced ice loss, *Earth Planet. Sci. Lett.* **289**(1), 22–29 (2010)
- 37.170 E.R. Ivins, T.S. James: Simple models for late Holocene and present-day Patagonian glacier fluctuations and predictions of a geodetically detectable isostatic response, *Geophys. J. Int.* **138**(3), 601–624 (1999)
- 37.171 E.R. Ivins, T.S. James: Bedrock response to Llanquihue Holocene and present-day glaciation in southernmost South America, *Geophys. Res. Lett.* **131**(L24613), 1–4 (2004)
- 37.172 S.A. Khan, K.H. Kjær, M. Bevis, J.L. Bamber, J. Wahr, K.K. Kjeldsen, A.A. Bjørk, N.J. Korsgaard, L.A. Stearns, M.R. van den Broeke, L. Liu, N.K. Larsen, I.S. Muresan: Sustained mass loss of the northeast Greenland ice sheet triggered by regional warming, *Nat. Clim. Change* **4**(4), 292–299 (2014)
- 37.173 Y. Jiang, T.H. Dixon, S. Wdowinski: Accelerating uplift in the North Atlantic region as an indicator of ice loss, *Nat. Geosci.* **3**(6), 404–407 (2010)
- 37.174 R. Dietrich, K. Rülke, J. Ihde, K. Lindner, H. Miller, W. Niemeier, H.-W. Schenke, G. Seeber: Plate kinematics and deformation status of the Antarctic peninsula based on GPS, *Glob. Planet. Change* **42**(1), 313–321 (2004)
- 37.175 I.D. Thomas, M.A. King, M.J. Bentley, P.L. Whitehouse, N.T. Penna, S.D.P. Williams, R.E.M. Riva, D.A. Lavallee, P.J. Clarke, E.C. King, R.C.A. Hindmarsh, H. Koivula: Widespread low rates of Antarctic glacial isostatic adjustment revealed by GPS observations, *Geophys. Res. Lett.* **38**(L22302), 1–6 (2011)
- 37.176 E. Rignot, G. Casassa, P. Gogineni, W. Krabill, A. Rivera, R. Thomas: Accelerated ice discharge from the Antarctic peninsula following the collapse of Larsen B ice shelf, *Geophys. Res. Lett.* **31**(L18401), 1–4 (2004)
- 37.177 T.A. Scambos, J.A. Bohlander, C.A. Shuman, P. Skvarca: Glacier acceleration and thinning after ice shelf collapse in the Larsen B embayment, Antarctica, *Geophys. Res. Lett.* **31**(L18402), 1–4 (2004)
- 37.178 G.A. Nield, V.R. Barletta, A. Bordon, M.A. King, P.L. Whitehouse, P.J. Clarke, E. Domack, T.A. Scambos, E. Berthier: Rapid bedrock uplift in the Antarctic peninsula explained by viscoelastic response to recent ice unloading, *Earth Planet. Sci. Lett.* **397**, 32–41 (2014)
- 37.179 P.L. Whitehouse, M.J. Bentley, A.M. Le Brocq: A deglacial model for Antarctica: Geological constraints and glaciological modelling as a basis for a new model of Antarctic glacial isostatic adjustment, *Quaternary Sci. Rev.* **32**, 1–24 (2012)
- 37.180 P.L. Whitehouse, M.J. Bentley, G.A. Milne, M.A. King, I.D. Thomas: A new glacial isostatic adjustment model for Antarctica: Calibrated and tested using observations of relative sea-level change and present-day uplift rates, *Geophys. J. Int.* **190**(3), 1464–1482 (2012)
- 37.181 E.R. Ivins, T.S. James, J. Wahr, O. Schrama, J. Ernst, F.W. Landerer, K.M. Simon: Antarctic contribution to sea level rise observed by GRACE with improved GIA correction, *J. Geophys. Res. Solid Earth* **118**(6), 3126–3141 (2013)

GNSS Part G

Part G GNSS Remote Sensing and Timing

38 Monitoring of the Neutral Atmosphere

Gunnar Elgered, Onsala, Sweden
Jens Wickert, Potsdam, Germany

39 Ionosphere Monitoring

Norbert Jakowski, Neustrelitz, Germany

40 Reflectometry

Antonio Rius, Cerdanyola del Valles, Spain
Estel Cardellach, Cerdanyola del Valles,
Spain

41 GNSS Time and Frequency Transfer

Pascale Defraigne, Brussels, Belgium

38. Monitoring of the Neutral Atmosphere

Gunnar Elgered, Jens Wickert

Global navigation satellite system (GNSS)-based atmosphere sounding techniques have become a widely recognized and operationally used remote sensing tool. A major milestone of this development was the beginning of the continuous use of GNSS data for improving regional and global forecasts in 2006. The principle behind these techniques is the utilization of atmospheric propagation effects on the GNSS signals on their way from the navigation satellites to receivers on the ground or aboard satellites. The atmosphere delays the time of arrival and introduces a curvature of the signal path. These effects can be accurately estimated and be used for the monitoring of the atmospheric variability. There are two different observation geometries. Therefore, we focus in the first part of this chapter on ground-based networks which are used to estimate the amount of water vapor above each receiver site. The second part deals with the use of radio occultation measurements from GNSS receivers aboard low Earth orbit satellites for global atmosphere sounding. We introduce and describe both techniques which provide observations suitable for the short-term weather forecasting and the long-term time series for climate research and monitoring.

38.1	Ground-Based Monitoring of the Neutral Atmosphere	1110
38.1.1	Accuracy of Propagation Delays	1111
38.1.2	From Delays to Water Vapor Content	1112
38.1.3	Applications to Weather Forecasting	1115
38.1.4	Applications to Climate Research	1118
38.2	GNSS Radio Occultation Measurements	1120
38.2.1	Introduction and History	1120
38.2.2	Basic Principles and Data Analysis	1120
38.2.3	Occultation Missions	1124
38.2.4	Occultation Number and Global Distribution	1125
38.2.5	Measurement Accuracy	1126
38.2.6	Prospects of New Navigation Satellite Systems	1127
38.2.7	Weather Prediction	1128
38.2.8	Climate Monitoring	1128
38.2.9	Synergy of GNSS Radio Occultation with Reflectometry	1131
38.3	Outlook	1132
	References	1133

This chapter deals with two distinctly different geometries: observations using ground-based global navigation satellite system (GNSS) networks and occultation observations from low Earth orbit (LEO) satellites. These are illustrated by the sketches in Fig. 38.1. In both geometries, the refractivity along the propagating path is determined by the atmospheric properties mainly in terms of pressure, temperature, and humidity.

The main application of the ground-based geometry is to infer the water vapor content above each receiver site on the ground. In principle, all the water vapor can be found within the troposphere, ranging from the ground up to 8–15 km. With a reasonable view of the

sky, there will always be a sufficient number of satellites visible in order to have continuous time series of the estimated water vapor content.

The radio occultation (RO) geometry is more dynamic, since here both the transmitting GNSS satellites and the receiving LEO satellites are in continuous motion with respect to the atmosphere. When occultations occur, height profiles of the refractive index in both the troposphere and the stratosphere are retrieved.

Because of very different geometries, these two methods use different processing techniques and different algorithms in the data analyses. They also produce completely different data products. Therefore, we first

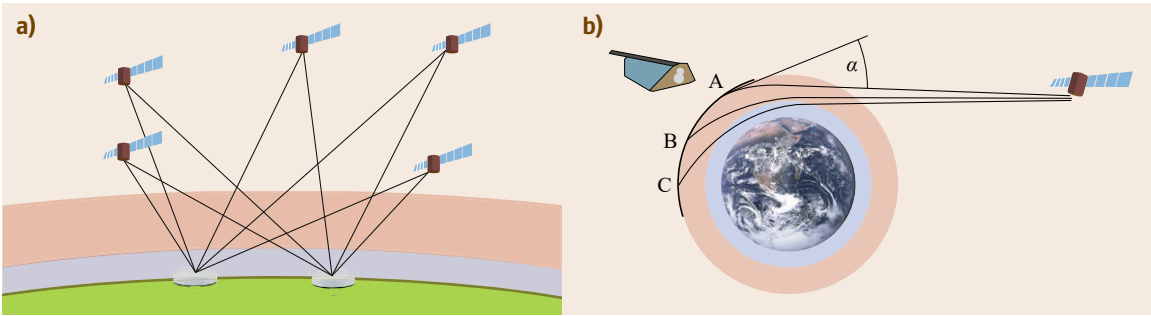


Fig. 38.1a,b Example geometries for ground-based observations (a) and ROs (b) of signals from GNSS satellites in a typical medium Earth orbit (MEO). The neutral atmosphere and the ionosphere are indicated by the blue and the red layers, respectively

discuss them separately. Applications of the remote sensing of the neutral atmosphere based on observations with ground-based networks are discussed in Sect. 38.1, while applications of GNSS receivers in LEO satellites are addressed in Sect. 38.2. The two geometries are of

complementary nature, and their strengths for applications in forecasting and research related to atmospheric processes over different timescales, from turbulence phenomena to climate issues, are finally summarized in Sect. 38.3.

38.1 Ground-Based Monitoring of the Neutral Atmosphere

As described in Chap. 6, a GNSS signal from a satellite is delayed in the atmosphere compared to propagation in vacuum. The effect, often described as an excess propagation path, is estimated in the GNSS data processing. This means that using a receiver on the ground, it is possible to infer the integrated amount of water vapor (IWV) in the atmosphere.

Figure 38.2 summarize the possible use of GNSS data in three different types of applications. First, the GNSS data can be used as a standalone product, for example, for monitoring the IWV at a specific site over long time. Second, the GNSS data can be used to assess, or verify, the results from numerical weather models used in forecasting or climate research. Third, the GNSS data be combined with other data in order to increase the quality, for example, when assimilated into

a numerical weather model in near-real-time weather forecasting.

In some applications, it can be a strength to estimate and to work with an integrated quantity, but, of course, for many other applications, profile information is necessary. For the ground-based geometry, the advantage is the accurate estimates of the IWV. This section focuses on this application, although attempts to retrieve profile information using tomographic methods is also discussed.

The radio-based space geodetic techniques of very long baseline interferometry (VLBI) and GNSS are affected by the atmosphere in terms of variations in the refractivity in the atmosphere which delays the time of arrival, the fundamental observable. It was shown in [38.1] and [38.2] that the estimated propagation delays from global positioning system (GPS) data, together with ground pressure observations, resulted in time series of the delays induced by the atmospheric water vapor, and they were in agreement with independent ground-based measurements from microwave radiometry.

The necessary background material needed in order to describe the estimation of propagation delays affecting signals, penetrating the atmosphere on their way to a receiver on the ground, is presented in Chap. 6. Let us here just repeat the basic definitions. The elevation dependence of the propagation delays is modeled by mapping functions, one for the hydrostatic delay and

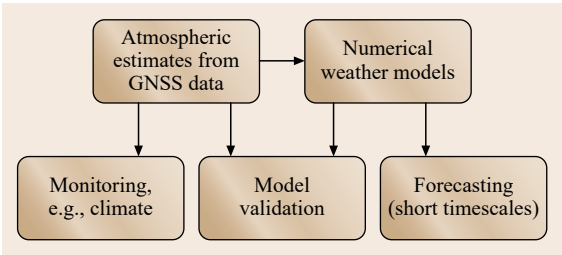


Fig. 38.2 Block diagram illustrating possible applications of atmospheric estimates from GNSS data in meteorology and atmospheric research

one for the wet delay in the geodetic GNSS data processing. The end result is the equivalent zenith total delay (ZTD). The ZTD

$$Z_t = Z_h + Z_w \quad (38.1)$$

is hence the sum of the zenith hydrostatic delay (ZHD)

$$Z_h = 10^{-6} \int_{h_0}^{h_\infty} N_h(z) dz \quad (38.2)$$

and the zenith wet delay (ZWD)

$$Z_w = 10^{-6} \int_{h_0}^{h_\infty} N_w(z) dz, \quad (38.3)$$

which are obtained from the integration of the hydrostatic (N_h) and wet (N_w) refractivities along the vertical propagation path from the height h_0 of the receiver to a point h_∞ outside the atmosphere.

With a sufficient number of observations in different directions during a defined time period, a refinement of just estimating an equivalent zenith propagation delay is possible. By defining linear horizontal gradients in the hydrostatic and wet refractivity profiles, an integrated parameter, normally referred to as the horizontal gradient, can be inferred. It was shown in [38.3] that by estimating horizontal gradients in the processing of GNSS data, the geodetic result improves significantly. It is today a common practice to estimate these gradients, although the use in meteorological applications has so far not been extensive. This potential will be further discussed later.

Another concept that occurs in the modeling of the neutral atmosphere is that of slant path delays. A simple method is to add the equivalent zenith delay and the estimated linear horizontal gradient and thereby model a specific delay in any given direction. The idea of adding, in addition, the residual delay toward each satellite from the GNSS data processing has also been proposed but shown to be wrong in the sense that systematic errors are introduced [38.4]. Instead, additional atmospheric parameters have to be estimated simultaneously, for example, by introducing tomographic methods.

38.1.1 Accuracy of Propagation Delays

Before describing the different applications of ground-based GNSS meteorology, we will review the error sources that determine the quality of the input data, that is, the estimated ZTD as described in Chap. 6.

The uncertainty of these excess propagation delays depends on several effects. Here, we review the relative importance of the type of mapping function used, mis-modeling of ionospheric effects (Chaps. 6 and 39), and effects caused by antennas and signal multipath.

Mapping Functions

The mapping functions define the elevation dependence of the hydrostatic and the wet delays, and are, therefore, an important parameter when estimating the atmospheric effect in the data processing. The development of more and more accurate mapping functions is described in Sect. 6.2.4. Many qualitative comparisons of mapping functions have been carried out over the years. Most mapping functions perform very well at high elevation angles, so when analyzing the accuracy of mapping functions, it is most interesting to focus on the results that are obtained at the lowest elevation angles, where the accuracy of the functions is decreasing. This means that, for each application, there is an optimum elevation cutoff angle. Low-elevation observations improve the geometry and reduce the formal error of the ZTD, but, at the same time, they introduce larger mapping function errors.

The *Niell* mapping function (NMF, [38.5]) has often been used due to its simplicity. It does not require any additional meteorological data; it only requires the location of the site and the time of the year. The size of uncertainties involved can be assessed by comparing the estimated ZTD using different mapping functions. As an example, a mean reduction of the ZTD of -2.6 mm was observed for 12 sites in Antarctica when changing from the NMF to the Vienna mapping function (VMF1) using elevation angles down to 7° [38.6].

Depending on the application, the need to include observations at low-elevation angles will vary. This is discussed later and now we just note that the choice of the mapping function is more or less irrelevant for GNSS meteorology if it is going to be used with observations acquired at elevation angles above say $15\text{--}20^\circ$.

Mismodeling of Ionospheric Effects

The accuracy of the ionospheric model used in the GNSS data processing affects the accuracy of the ZTD. Errors in the modeling of the signal delays caused by the ionosphere are more or less absorbed into the delays estimated for the neutral atmosphere, that is, the ZTD. To the authors' knowledge, no study has been directly focused on the accuracy of the estimated ZTD for different methods to correct for the ionospheric influence. However, there have been studies addressing the accuracy of the estimated position for different methods of handling the ionosphere. There is a strong correlation between the estimated vertical coordinate and

the estimated ZTD. A relative error in the ZTD is approximately three times smaller than the error in the estimated vertical position. The factor depends on the geometry, in terms of the elevation cutoff angle used for the observations [38.7]. Therefore, studies on the influence of the ionospheric model on the estimated site position indirectly provide information on the accuracy in the estimated ZTD.

The inclusion of higher order terms in the ionospheric model showed a systematic effect at the level of several millimeters in the station positions [38.8]. Later a more accurate model of the International Geomagnetic Reference Field (IGRF) was used [38.9]. Different models for the geomagnetic field have also been studied in [38.10], resulting in the recommendation that corrections for higher order ionospheric effects shall be included, particularly in equatorial regions and over periods of solar maximum when the ionosphere is more active. Given the 11-year-long solar cycle, it is reasonable to assume that the recommendations from such studies will depend on the timescale of the application.

Antenna Phase-Center Variations

Antenna phase-center variations exist both at the satellite antenna and at the receiving antenna on the ground. It has been shown that phase center variations (PCVs) (Chap. 17) have a significant influence on the estimated delay in the neutral atmosphere [38.11], and hence also on the IWV. The variation depends on the nadir angle from the satellite to the ground or, equivalently, the elevation angle of the satellite seen from the ground. The main effect is a bias-type of error but will, of course, change with changing geometries. Recommendations for modeling antenna PCVs exist (Chaps. 19 and 25) and shall be used in order to reduce their influence. Proper modeling of PCVs is especially relevant for

climate-related applications investigating (small) long-term trends in the IWV. An assessment of the effect caused by the introduction of new GPS satellite types with different antenna phase patterns revealed an artificial trend of up to roughly $0.15 \text{ kg}/(\text{m}^2 \text{ year})$ for the estimated IWV over a five-year period [38.12].

Signal Multipath

Signal multipath (Chap. 15) degrades the precision of the arrival time of the signal from the satellite. The effect will have a more or less strong dependence on the elevation angle of the direct signal, depending on the local electromagnetic environment at the receiving antenna. It is difficult to model since the environment is changing. For example, the reflective properties of the ground change when it is covered by (rain) water, snow, and soil moisture, if soil is present. It has been shown that the effect of signal multipath can be reduced by mounting a plate with a microwave-absorption material just below the GNSS antenna [38.13].

38.1.2 From Delays to Water Vapor Content

The primary estimate from the data processing for the atmospheric influence is the ZTD. For some meteorological applications, the ZTD can be used directly, whereas other applications require the time series of the IWV, for example, when validating results from other instruments where the IWV is the primary output.

The overall data flow is illustrated in Fig. 38.3. The main operation is first to subtract the ZHD from the ZTD in order to obtain the ZWD. Thereafter the IWV is calculated from the ZWD. It shall be pointed out that the use of numerical weather models is not necessary. It is, however, common that information from such models is used to derive and optimize mapping functions,

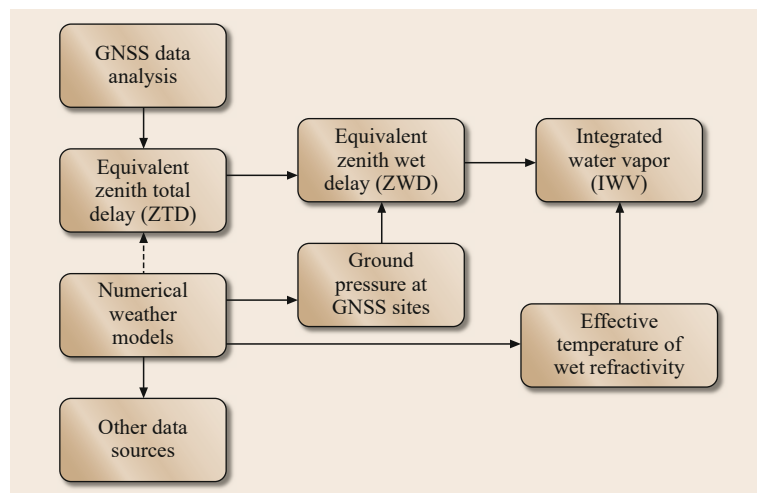


Fig. 38.3 The data flow for different applications in ground-based GNSS meteorology. Depending on the application the timescales of the data flow between the different operations can vary from seconds to years

and they may also provide a priori information about the atmosphere to the GNSS data analysis. Since the mapping functions for the hydrostatic and the wet delays are different such a priori information has been shown to increase the accuracy of the estimated ZTD [38.6]. As will be discussed below, there are also alternatives to numerical weather models for the estimation of the ground pressure and the effective temperature of the wet refractivity.

Before discussing these further, we note that both steps will require some knowledge of the state of the atmosphere, mainly the profiles of pressure and temperature. For applications such as weather forecasting, up-to-date knowledge already exists in the numerical weather models, and, in such cases, there is no need to compromise by using less accurate relations. Instead, the normal procedure is then to assimilate the ZTD directly into the numerical weather model.

From ZTD to ZWD

The first step is to subtract the ZHD from the ZTD. Hence, the uncertainty in the ZHD is directly transferred to the ZWD via (38.1) and (6.50). The only observable needed for the calculation of the ZHD is the ground pressure, assuming that the latitude and the height of the station are approximately known. This assumption is indisputable for ground-based GNSS networks used for meteorological applications. There are basically two possibilities to obtain the ground pressure: either to use a barometer at the site or to use the analysis from a numerical weather model which has input observations of the ground pressure of a sufficient accuracy from the surrounding area.

Commercially available pressure sensors provide accuracies much better than 0.5 hPa. Figure 38.4 depicts the observed differences between three barometers at the Onsala VLBI site. Comparison with a barometer from the Swedish Meteorological and Hydrological

Institute (SMHI) shows that the present sensor has an absolute accuracy at the level of better than 0.2 hPa. Over the time period of interest, the SMHI barometer has been calibrated approximately every second year, and it is traceable to the SI unit within 0.1 hPa.

Several studies assessing the accuracy of deriving the ground pressure at the site from numerical weather models have been performed. The uncertainty of the ground pressure derived from the European Centre for Medium-Range Weather Forecasts (ECMWF) has been evaluated [38.16]. They compared the interpolated ground pressure from the ECMWF analysis to the local ground measurements at more than 60 globally distributed GPS sites using one year of data. The results revealed an agreement with an overall mean bias and a standard deviation of 0.0 hPa and 0.9 hPa, respectively. A similar test, using more than 10 years of data for the GPS site at the Onsala Space Observatory, resulted in a mean bias and a standard deviation of 0.1 hPa and 0.6 hPa, respectively [38.17]. Pressure sensors of the World Meteorological Organization (WMO) were compared to nearby (< 50 km) locally installed sensors and biases of less than 1 hPa for more than 90% of the stations were found [38.18]. Similar results were obtained in [38.19] using independent data sets and models.

These uncertainties together with other parameters in (6.50) are summarized in Table 38.1. Assuming uncorrelated errors, these four add up to a relative uncertainty equal to $2.2 \cdot 10^{-4}$. To this uncertainty shall also be added the uncertainty in the ground pressure. This is illustrated in Fig. 38.5, where the ground pressure uncertainties of 0.2 hPa and 1.0 hPa are assumed to be relevant for the observations and models, respectively. We conclude that the major contribution to the total ZHD uncertainty is from model uncertainties of 1.0 hPa, equivalent to a relative error of $1 \cdot 10^{-3}$, which corresponds to ≈ 2 mm in the ZHD. If actual high-quality pressure observations are used instead, the

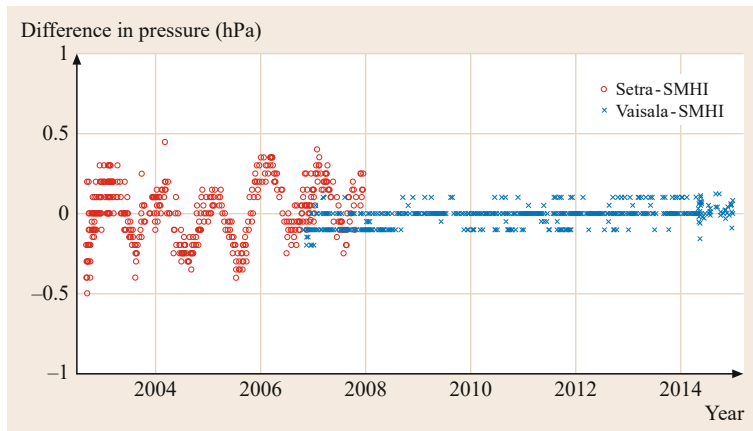


Fig. 38.4 Differences in the observed pressure by three different barometers. The Setra barometer was the standard unit at geodetic VLBI stations when the Mark III system was introduced. The small annual variations are likely caused by a sensor sensitivity to temperature. These are completely removed when the new barometer from Vaisala was connected to the VLBI system

Table 38.1 Parameters used to calculate the ZHD, their uncertainties, and the resulting uncertainties in the ZHD (Z_h)

Parameter	Value	Uncertainty	Unit	Relative uncertainty	Reference
k_1	77.6890	0.015	K/hPa	$1.9 \cdot 10^{-4}$	Table 6.2
R	8.3144621	0.0000075	$\text{J mol}^{-1} \text{K}^{-1}$	$9.0 \cdot 10^{-7}$	[38.14]
M_d	28.9644	0.0014	kg kmol^{-1}	$4.8 \cdot 10^{-5}$	[38.15]
g_{eff}	≈ 9.784	0.001	m s^{-2}	$1.0 \cdot 10^{-4}$	(6.51), [38.15]
p_0 , case 1	≈ 1000	0.2	hPa	$0.2 \cdot 10^{-3}$	Typical observation, see the text
p_0 , case 2	≈ 1000	1.0	hPa	$1.0 \cdot 10^{-3}$	Typical model, see the text
Z_h , case 1	≈ 2.28	$0.7 \cdot 10^{-3}$	m	$0.3 \cdot 10^{-3}$	Typical observation, see the text
Z_h , case 2	≈ 2.28	$2.3 \cdot 10^{-3}$	m	$1.0 \cdot 10^{-3}$	Typical model, see the text

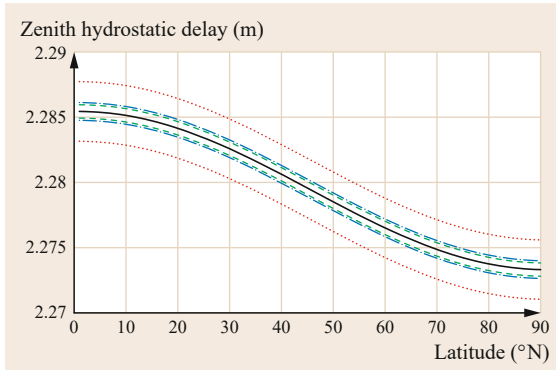


Fig. 38.5 The expected ZHD when a ground pressure of 1000 hPa is observed from the sea level (solid black line). The uncertainty in the ZHD caused by the uncertainties in the constants used in the conversion is indicated by the green dashed lines. The total uncertainty, when errors in the pressure observations of 0.2 hPa and 1.0 hPa are taken into account, is shown by the blue dash-dotted and the red dotted lines, respectively. All errors are assumed to be uncorrelated and added as root-sum-squared

uncertainty from the value of k_1 becomes equally important, resulting in a relative uncertainty of $0.3 \cdot 10^{-3}$. This means an uncertainty in the ZHD of less than 1 mm.

Finally, before leaving the subject of calculating the ZHD, it shall be noted that one additional uncertainty exists: the approximation of hydrostatic equilibrium. This approximation is good to one part in 10^4 [38.20]. The effect has been studied at two specific laser-ranging sites in mountainous areas. It was concluded that deviations from hydrostatic equilibrium may cause errors in excess of 1 cm, corresponding to 3 mm in the zenith direction, although it is an unlikely event [38.21].

From ZWD to IWV

The water vapor content, V , is defined as

$$V = \int_0^{\infty} \rho_v dh, \quad (38.4)$$

where ρ_v is the absolute humidity in g/m^3 and h is the height in m. An alternative parameter often used to describe the water vapor content of the atmosphere is the precipitable water (PW). This is a measure of the equivalent height of the column formed if all the water vapor is condensed and collected at the ground surface, that is, a PW value of 1 mm is equivalent to an IWV value of 1 kg/m^2 .

Using the ideal gas law, we can instead use the partial pressure of water vapor e and the temperature T and obtain

$$V = \frac{1}{\rho_w R_w} \int_0^{\infty} \frac{e(h)}{T(h)} dh, \quad (38.5)$$

where ρ_w is the density of liquid water and R_w is the specific gas constant of water vapor.

We note that the expression for the ZWD is similar to

$$Z_w = 10^{-6} \left(k'_2 \int_0^{\infty} \frac{e(h)}{T(h)} dh + k_3 \int_0^{\infty} \frac{e(h)}{T(h)^2} dh \right), \quad (38.6)$$

where k'_2 and k_3 are the constants determined from laboratory experiments of the refractivity. The values are given in Table 6.2 and (6.34) of Chap. 6.

We can combine (38.5) and (38.6) and express the IWV in the ZWD as

$$V = \frac{Z_w}{Q} \quad (38.7)$$

where

$$Q = 10^{-6} \rho_w R_w \left(\frac{k_3}{T_m} + k'_2 \right). \quad (38.8)$$

The parameter T_m can be estimated from the vertical profiles of the atmospheric temperature and the partial

pressure of water vapor

$$T_m = \frac{\int_0^\infty T(h) \frac{e(h)}{T(h)^2} dh}{\int_0^\infty \frac{e(h)}{T(h)^2} dh} \quad (38.9)$$

It can be seen as a mean atmospheric temperature weighted by (e/T^2) . Now, let us study the uncertainties introduced when calculating the IWV using the ZWD as an input. The relation between these two parameters is defined by (38.7), involving the density of wet air, the specific gas constant for wet air, k'_2 , k_3 , and T_m . We note that T_m is the only one of these which will vary spatially and temporally.

In order to obtain T_m for a specific GNSS site, an obvious and accurate method is to use the vertical profiles of atmospheric temperature and humidity. Such profiles may be obtained from the reanalysis based on a numerical weather model. If such tools are not available, one may use a more simple relation. A linear relation was derived in [38.22]

$$T_m = 70.2 + 0.72 T_s, \quad (38.10)$$

where T_s is the surface temperature in K.

A global study, using six years of data, of the uncertainty in T_m calculated from the surface temperature used results from numerical weather models and radiosonde observations [38.23]. It showed that the root mean square (rms) error was dominated by a mean bias. For example, the bias found when using (38.10) varied in the interval $\pm 3.5\%$. When the bias is removed, the remaining error was less than 0.5% over most of the globe.

We can also calculate values of Q using a model optimized for a specific region based on the annual variability in Q due to seasonal temperature variations. For Europe, such a model

$$Q = a_0 + a_1 \theta + a_2 \sin\left(2\pi \frac{t_D}{365}\right) + a_3 \cos\left(2\pi \frac{t_D}{365}\right) \quad (38.11)$$

was derived from radiosonde data [38.24]. Here, θ is the site latitude in degrees and t_D is the decimal day of the year. Values for the coefficients a_0 , a_1 , a_2 , and a_3 are given in [38.25] based on radiosonde data from 38 sites in Europe and spanning a period from 1989 to 1997. The resulting rms error in the IWV is of the order of 1.5%.

To conclude the discussion on how to estimate T_m , there can be calculations of statistical averages for sites, regions, or the whole globe. For a specific region or for a specific site, it is possible to develop a model including yearly variability which is often based on observational data, for example, from radiosondes. A more accurate method is to use an analysis based on a numerical weather model. There are several similarities to the technique of how to optimize the mapping function for a given site/region. The issue will be further discussed in the section on climate applications.

We can now assess the relative importance of the different contributions to the total uncertainty in the IWV. Summarizing the discussions above, we have:

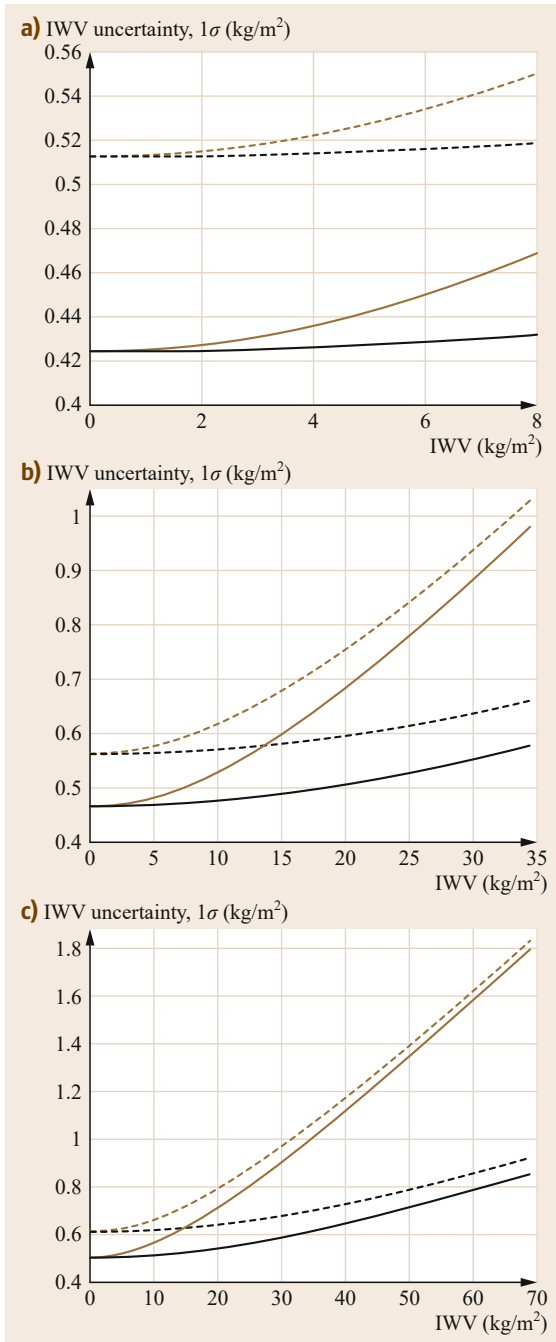
- An uncertainty of 2–5 mm in the ZTD from the processing of GNSS data.
- An uncertainty in the ZHD, which is determined by the uncertainties of the parameters in Table 38.1 and illustrated in Fig. 38.5.
- The uncertainties introduced by the conversion from ZWD to IWV. These include uncertainties in k'_2 , k_3 , the specific gas constant for water vapor R_w , and the mean temperature T_m . The temperature dependence of the density of liquid water is sufficiently small over the range of atmospheric temperature so that it can be neglected [38.26].

In order to compare the relative importance of all these contributions, a number of different assumptions about these uncertainties are made in Fig. 38.6. These plots show the uncertainties for a dry and cold, a temperate, and a hot and humid troposphere. We note that in a dry and cold troposphere, the IWV uncertainty is mainly caused by the assumed accuracy of 3 mm in the ZTD plus the assumed accuracy in the ground pressure observations. As the troposphere becomes warmer and more humid, the relative importance of the uncertainty in the conversion factor Q increases.

38.1.3 Applications to Weather Forecasting

Water vapor is an important parameter determining the state of the atmosphere. A general understanding can be obtained by studying the water cycle in the atmosphere. In short, it can be described as follows:

Water on the ground – in the oceans, lakes, streams, and vegetation – evaporates and transpires into the atmosphere. It carries energy, which is released when the water vapor condenses into clouds. These clouds may form precipitation more or less immediately, or at a later stage when the atmospheric conditions cause the liquid drops formed to be large enough to fall back to the surface of the Earth.



Knowledge about the amount of water vapor in the atmosphere is mainly important for short-term forecasts or nowcasting. The IWV is highly variable in space and time. For example, it may change by a factor of 2 in just a couple of hours due to moving mesoscale weather systems carrying different types of air masses in terms of temperatures and humidities. In order to be useful in

Fig. 38.6a–c The expected uncertainty in terms of one standard deviation in the IWV. Three different weather conditions are presented: cold ($T_m = 250 \text{ K}$, $Q = 7.26$, (a)), medium ($T_m = 275 \text{ K}$, $Q = 6.63$, (b)), and hot ($T_m = 300 \text{ K}$, $Q = 6.10$, (c)). Since the temperatures are strongly correlated with the absolute humidity in the atmosphere the chosen ranges for the IWV are different for different cases. In all cases we assume an uncertainty in the ZTD of 3 mm. We add an uncertainty due to the hydrostatic delay for two different uncertainties in the ground pressure: either 0.2 hPa (solid lines) or 1.0 hPa (dotted lines) and finally the uncertainty from the conversion from ZWD to IWV for two different uncertainties in the parameter Q : either 1.0% (black lines) or 2.5% (light brown lines). All errors are assumed to be one standard deviation, uncorrelated, and are added as root-sum-squared ◀

weather forecasting, the IWV results must be available within a couple of hours, but the sooner the better.

Because warm and cold air masses often correlate strongly with the IWV, large-scale motions of mesoscale systems are easily tracked by ground-based networks. Distinct cold or warm fronts can be the cause of significant spatial and temporal gradients in the IWV above a GNSS site. Combining time series from many sites makes it possible to track such weather systems [38.27]. Such a spatiotemporal structure was assessed by an Empirical Orthogonal Function (EOF) analysis, where over 90% of the water vapor variability is explained using the first temporal eigenvector only [38.28]. For very dense networks and weather situations that are accurately described by a *frozen flow*, it has been shown that estimates of wind speed and wind direction can be made [38.29]. It shall, of course, be noted that this is an indirect method, and when the frozen flow hypothesis is not valid, the method will break down.

Large investments have been made in continuously operating reference networks for surveying and real-time kinematic (RTK) applications. Such networks are typically established on a national scale by government bodies, but also commercial networks exist. This means that not all data are openly available in real time which in practice means some restrictions on how data can be distributed for close to real time processing. An example is the EUMETNET GNSS Water Vapour Programme (E-GVAP). A snapshot of the status at a specific time is shown in Fig. 38.7. (EUMETNET is a group of European National Meteorological Services that provides a framework for co-operative programmes in the various fields of basic meteorological activities.)

In Germany, where many stations participate in the E-GVAP network, the receivers are distributed with

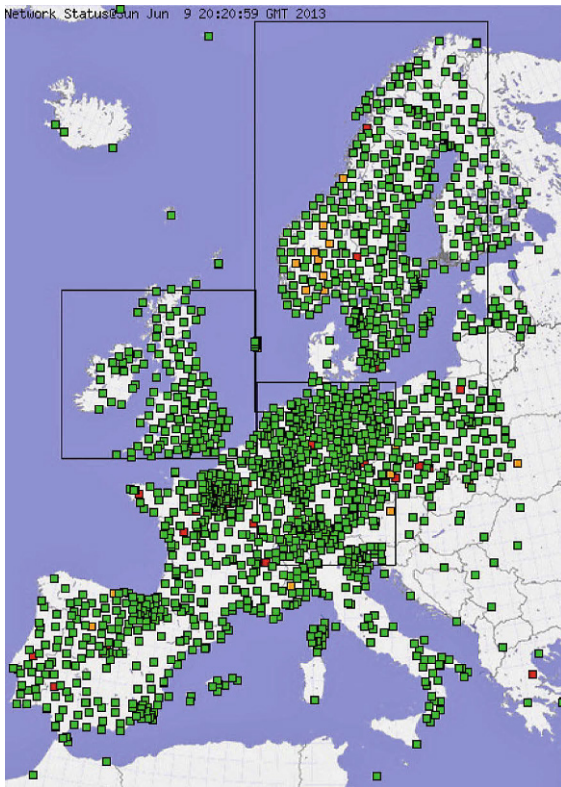


Fig. 38.7 The E-GVAP project is an example of a network of GNSS receivers consisting of several subnetworks where data processing is distributed to several centers. This example is from 2013 and the number of stations continues to increase (after [38.30])

a rather high spatial resolution and baseline lengths of the order of 40 km. An example of GNSS-based results of the IWV is shown in Fig. 38.8.

Assimilation of the estimated IWV was first considered when using GNSS data for weather forecasting. However, it was almost immediately realized that because the necessary information, in terms of pressure and temperature fields, is already available in the numerical weather model it is an advantage to instead assimilate the ZTD.

The method of three-dimensional variational data assimilation (3D-Var) is often used with a typical update period of 3 h. However, the four-dimensional variational data assimilation (4D-Var) method offers to benefit from the much higher temporal resolution of the GNSS results. The timing of the passage of weather systems and different dry and wet air masses is crucial for short-term weather forecasts [38.31].

There have been many assessments of the impact on the quality of weather forecasts when using GNSS data. Assimilation of ZTDs has, for example, shown im-

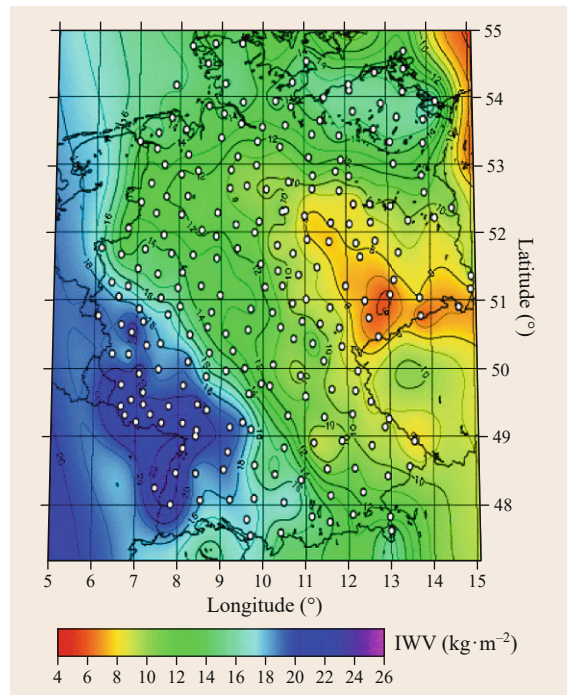


Fig. 38.8 The IWV over Germany on Feb. 28, 2010, 00:07 coordinated universal time (UTC), derived from ground-based GNSS stations. The white circles denote the locations of GNSS receiver sites (courtesy of G. Dick, GFZ)

provements in precipitation and cloud cover forecasts (see [38.32] and [38.33], respectively).

A further refinement may be to assimilate slant delays [38.34]. These delays may be estimated by combining the ZTD with estimated linear gradients. An alternative in order to study small-scale variations in the atmospheric water vapor (without assimilation) is to apply tomographic methods to very dense ground-based networks. Having all GNSS receiving antennas on the ground will, however, imply a weak geometry for the inversion algorithms used in tomography. This can to some extent be compensated for by introducing constraints on the variability in the water vapor density between the different volume pixels defined and used in the estimation method [38.35, 36]. The geometry is, of course, improved if the receivers are located in a landscape where the height differences are large as this will directly yield differential IWV values for different atmospheric layers. Such studies have, for example, been performed on Hawaii [38.37].

In the future, one can imagine that the raw GNSS observations are assimilated, effectively meaning that the entire GNSS data processing is executed in the numerical weather model.

38.1.4 Applications to Climate Research

Water vapor in the atmosphere is of great relevance also for climate research because it is a very important parameter in the water cycle as well as an efficient greenhouse gas. An increase of 20% of the IWV in the tropics has a larger impact than a doubling of the carbon dioxide concentration [38.38].

One key question is to quantify the positive feedback due to an increase in the IWV. A study of both the short-term and the long-term feedback concluded that the time series of observed water vapor need to be longer than 25 years in order to accurately determine the effect [38.39]. The GNSS ground-based networks established in the mid-1990s will, hence, have the potential to be useful for such applications in the 2020s.

Within the Global Climate Observing System (GCOS), there is a specific international reference observing network called the GCOS Reference Upper Air Network (GRUAN). One component in GRUAN is ground-based GNSS observations in order to provide IWV time series at selected reference sites where several independent observing techniques are available [38.40].

We first present some examples of GNSS results focusing on different timescales: trends over decades, annual components, and diurnal components. These may be used for climate monitoring and for evaluating of climate models which concludes this first part of GNSS meteorology using ground-based networks.

Long-Term Trends

There is a demand for long and stable time series for monitoring and therefore there is a need to assess the uncertainty at an absolute level, or at least as a stability measure, over decades. Since errors in the empirically determined constants k_1 , k_2 , and k_3 will not influence the uncertainties in the observed trends, for this application, it is more appropriate to refer to requirements on the long-term stability than on the absolute accuracy.

The difficulty to estimate trends that are of the order of a few percent over decades is illustrated in Fig. 38.9. The variability is huge – not only on a day-to-day basis but also over the seasons and from one year to another. Therefore, it is important to note that estimates of linear trends by no means shall be expected to be identical for adjacent periods of several years.

In addition to the large variability in weather, an additional issue is change (controlled or uncontrolled) in the electromagnetic environment of the receiving antenna. Examples of such changes are installations of different types of antennas and radomes, as well as their orientation. For the case of estimating a trend in the ver-

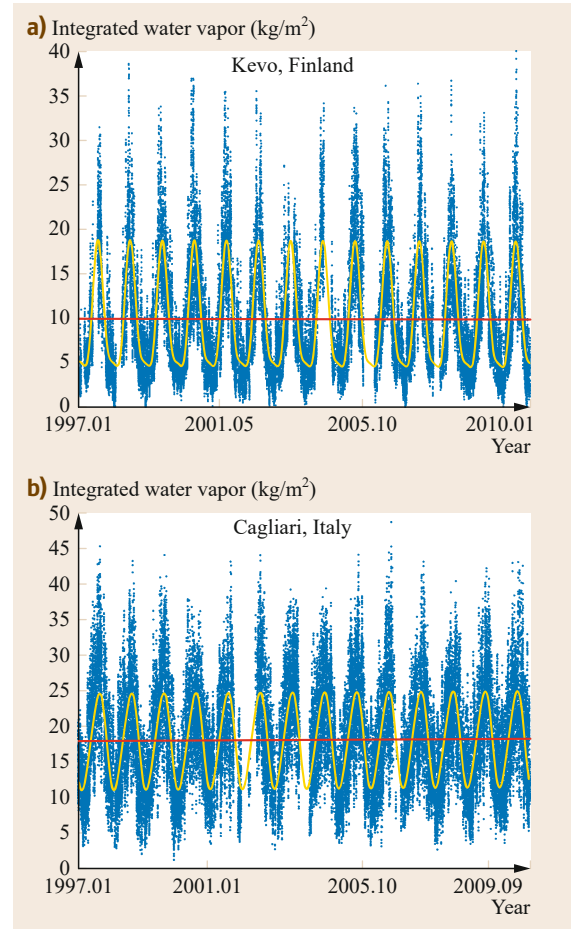


Fig. 38.9a,b Time series of the water vapor content at Kevo, Finland (**a**) and Cagliari, Italy (**b**). A model including a mean value plus a linear trend (red line) and a seasonal component (yellow line) is fitted to the hourly estimates (blue dots)

tical coordinate, when there are reasons to believe that the trend is constant over the length of the time series, one may choose to estimate an additional offset at the time of the intervention (see Fig. 38.10 and [38.41] for more details).

As already mentioned, it is not reasonable to assume that a true trend in the IWV should remain constant over many years, in spite of the fact that an estimated linear trend is an obvious parameter to estimate as an indicator of a change in the climate.

A similar method was assessed in [38.42], where an effort was made to model the impact of a change of radome at Onsala, Sweden, on February 1, 1999. It was more reliable in this case to study the mean difference of the wet delay between the GNSS and the VLBI techniques and to apply the observed change in offset

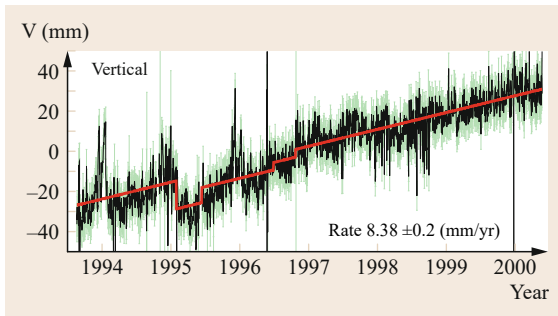


Fig. 38.10 Estimates of the vertical coordinate using GPS data from the station Sveg, Sweden (courtesy of Scherneck & Haas)

as a constraint when estimating a trend from the full time series, thereby indirectly using the VLBI technique as an independent data source to calibrate the absolute scale of the IWV from the GNSS data.

A possibility of handling false jumps in the IWV time series when no independent results are available could be to make use of the correlation between estimates of station coordinates and the propagation delay. This idea requires further studies.

Annual Components in the IWV

Annual components in the IWV show large differences over the globe caused by different weather patterns related to the seasons. Through continuously operating ground-based networks and a homogeneous processing, GNSS is a tool to continue such systematic studies. For example, 13 years of data from 155 globally distributed GNSS sites were studied in order to conclude that the large seasonal amplitudes occurred at mid latitudes [38.43].

Also, more local studies, with a higher spatial resolution, have been carried out. For example, a 10 year long period was studied using GNSS data from the Iberian peninsula in [38.44] and [38.45] revealed a systematic deviation in the southwest of Spain during the summers.

Diurnal Components in the IWV

The diurnal variability in the IWV is driven by the incoming solar radiation. Hence, it is more dominant in the equatorial region, and as we go away from the equator, the amplitude decreases and vanishes in the areas close to the poles. For the same reason, and because the absolute humidity typically increases with the temperature, it is also expected to be larger during the summer compared to the winter. The knowledge of its amplitude and phase can be a useful tool in order to assess the accuracy of numerical weather models used both for forecasting and in climate research.

In a global study using 1 year of data from 151 International GNSS Service (IGS) sites, diurnal amplitudes between 0.2 mm and 10.9 mm in the ZTD were observed [38.46]. This corresponds to approximately from less than 0.1 up to 1.7 kg/m². A similar study also using a US network [38.47] obtained similar results and also assessed the accuracy of three different reanalysis products.

In a study using 14 IGS sites in the equatorial region, diurnal amplitudes up to 3 kg/m² were observed [38.48]. The amplitudes were larger for sites that are not close to the sea, which is expected, in general, because of the lower diurnal variability of the temperature close to the sea compared to more inland areas.

It may be worth noting that the diurnal components at mid to high latitudes is hidden by the much larger moving mesoscale weather systems. This together with the fact that the diurnal variation of the solar radiation has a much smaller amplitude means that typically many years of GNSS data must be averaged in order just to detect the diurnal component [38.49].

Evaluation of Climate Models

We have already touched upon the application of using the GNSS data in order to evaluate numerical climate models for the trends, annual, and diurnal components in IWV time series.

An additional study that focused both on the validation of seasonal and interannual variations in the IWV was presented in [38.50]. The authors of this study found, in general, an agreement at the submillimeter level for the precipitable water in Europe and North America between GNSS and a numerical weather prediction model from National Centers for Environmental Prediction (NCEP). On the other hand, the model was found to underestimate the seasonal signal by up to 40% and 25% for the equatorial region and Antarctica, respectively.

In [38.51], a regional climate model was evaluated in terms of the IWV differences between GNSS observations and the climate model. It was found that a couple of GNSS sites showed large differences, which in turn were attributed to a cold temperature bias and an underestimate of the diurnal temperature range for the model in that area. It was also noted that the model produced a positive bias in the IWV at sites close to the sea (the surface tile of the model gridpoint has a water coverage > 60%), possibly due to the fact that evaporation in the model has a too high influence on the IWV mean value for the gridpoint.

As mentioned several times already, climate studies require long-term averages. It therefore seems appro-

priate to note that so far the potential of ground-based GNSS has been demonstrated, but there is a limitation today on the studies that are meaningful due to the length of available time series. This situation will however improve over the years to come.

The next section will deal with the occultation geometry in GNSS meteorology. Thereafter, we will conclude the chapter by summarizing the GNSS applications for the remote sensing of the neutral atmosphere.

38.2 GNSS Radio Occultation Measurements

38.2.1 Introduction and History

On July 17, 1995 the U.S. Air Force announced

[...] that the Global Positioning System satellite constellation has met all requirements for Full Operational Capability.

Already before this announcement on April 5, 1995 the LEO MicroLab-1 satellite was launched and recorded for the first time at all signals from setting GPS satellites, which tangentially traversed the Earth's atmosphere. The main purpose of these observations was atmosphere sounding using the innovative GPS RO technique within the GPS/MET(GPS/METeorology)-Experiment [38.52, 53].

GPS/MET was a real story of success. For the first time, globally distributed vertical profiles of atmospheric temperature, water vapor, and electron density were successfully derived from spaceborne GPS data. The GPS (or, more generally, GNSS) RO technique became reality as a new and innovative remote-sensing method. The properties of this calibration-free atmospheric limb-sounding technique (e.g., all-weather-capability, high accuracy, high vertical resolution, low-cost realization) promised to have a great potential for atmospheric and ionospheric research, numerical weather forecasts, space weather monitoring, and climate change detection [38.54].

Around 20 years later, it can be stated, that GNSS RO kept this promise and is widely recognized as an established atmospheric remote-sensing technique. A major and prominent example for this development is the beginning of the operational use of GNSS RO data to improve global numerical weather forecasts (e.g., [38.55, 56]). Figure 38.11 shows a schematic illustration of the GNSS observation geometry. A GNSS receiver aboard an LEO satellite tracks the signals (carrier-phase and amplitude) of an occulting GNSS satellite, that is, within the period directly before satellite *set* or before satellite *rise*. These are the occultation events and last typically 1–2 min for atmosphere sounding from the Earth's surface up to around 100 km. During these events, the signal goes through different vertical layers of the atmosphere and is modified in

a characteristic way. By appropriate inversion of the time series of the signals during the occultation, vertical profiles of atmospheric parameters, as refractive index, temperature, or water vapor can be derived. GNSS RO can also be used to derive vertical electron density profiles, as described in more detail in Chapt. 40 of this book. A key observable is the bending angle α of the signal path from the occulting GNSS to the LEO satellite, which is assigned to the impact parameter a and the point of the closest approach of the signal path to the Earth's surface r_0 . Additional LEO measurements from a referencing GNSS satellite and GNSS ground station data from the occulting and referencing satellite are used for the calibration of the atmospheric excess phase of the occultation measurements, which is the base for the bending angle derivation. More details are given in Sect. 38.2.2.

38.2.2 Basic Principles and Data Analysis

Derivation of the Atmospheric Excess Phase

The GNSS RO technique is based on precise dual-frequency (for ionosphere correction) phase measurements of a GNSS receiver in an LEO, which is tracking setting or rising GNSS satellites. Combining these measurements with the satellites' position and velocity information, the phase path increase due to the atmosphere during the occultation event can be derived. This phase path increase is called atmospheric phase delay or atmospheric excess phase, dA and the geodetic key observable of GNSS RO and its derivation here briefly reviewed.

The observed phase L for each frequency of the occultation link (see in Fig. 38.11 between occulting GNSS satellite and LEO) in units of meters can be written as

$$L = \rho + c(dt - dT) - dI + dA + \epsilon. \quad (38.12)$$

Here, ρ denotes the true range between the transmitter and receiver taking into account the signal travel time, c is the velocity of light, dt and dT are the transmitter and receiver clock errors, respectively, dI and dA are the phase delays due to ionosphere and neutral atmosphere

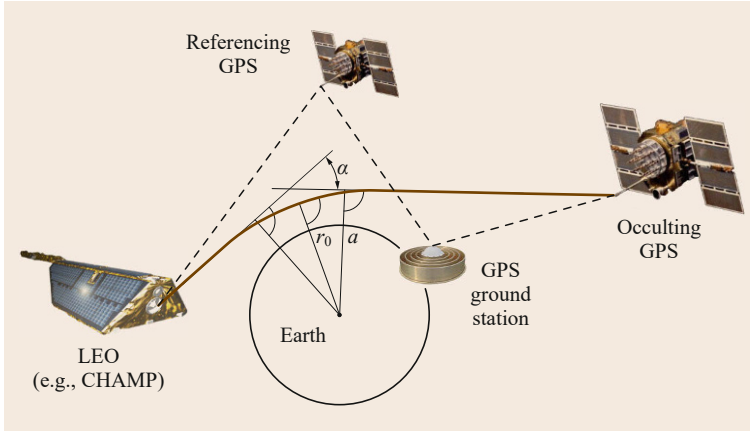


Fig. 38.11 The principle of GNSS RO measurements aboard a LEO satellite such as CHAMP (CHALLENGING Minisatellite Payload). A key observable is the atmospheric bending angle α of the signal path from the occulting GNSS to the LEO satellite. Under assumption of spherical symmetry an impact parameter a can be assigned. LEO measurements from a referencing GNSS satellite and GNSS ground station (*dashed lines*) are used for calibration of the RO measurements.

along the ray path, respectively, and ϵ is a residual error composed of, for example, measurement noise and uncorrected multipath.

For the analysis of the GPS/MET and the initial CHALLENGING Minisatellite Payload (CHAMP) measurements [38.52, 57], a double-difference technique was used to eliminate the GNSS transmitter and LEO receiver satellite clock errors: the signals from the occulting GNSS satellite were differenced with those from a reference GNSS satellite. These satellite measurements were synchronized with simultaneously recorded data provided by a fiducial ground network [38.58]. The corresponding observation geometry is depicted in Fig. 38.11.

The double difference,

$$\Delta\Delta L = (L_{CO} - L_{CR}) - (L_{GO} - L_{GR}), \quad (38.13)$$

is formed from simultaneous CHAMP and ground station measurements of signals from both the occulting and the referencing GNSS satellite during an occultation (e.g., [38.58]). The subscripts C, O, R, and G denote CHAMP, occulting and referencing GPS satellite, and the ground station, respectively. The corrections of relativistic and light time effects have to be taken into account [38.59]. Equation (38.13) shows that in the double-difference method, both the transmitter clock errors dt_O and receiver clock errors dt_R cancel. While the double-difference method eliminates the satellite clock errors, other errors are introduced by the three auxiliary satellite links involved. These errors are uncalibrated atmospheric and ionospheric contributions and additional noise. Furthermore, for differencing with nonsynchronous receiving times of occultation and reference satellite, the ground receiver clock drifts and also multipath wave propagation at the ground station location have to be taken into account (e.g., [38.60]).

Due to the termination of Selective Availability (SA) on May 2, 2000, which reduced the apparent variations in the GNSS transmitter clocks by various orders of magnitude, and due to the higher stability of presently available LEO satellite clocks, the application of single- and even zero-differencing analysis techniques represents the current state of the art for GNSS RO data analyses [38.61, 62].

For example, the space-based single difference,

$$\Delta L = (L_{CO} - L_{CR}) \quad (38.14)$$

is the difference between phase measurements of CHAMP from the occulting GNSS, L_{CO} , and the referencing GNSS, L_{CR} , respectively. In this scheme, the GNSS satellite clock errors remain and need to be corrected for, what is feasible after the termination of SA even with the standard data products of the IGS. For GNSS RO satellites with ultra stable oscillators (USOs), as, for example, GRACE-A or Metop, even the forming of single differences is not required and the phase data of the occultation link L_{CO} can be directly used for the derivation of dA . With single and zero differencing, the level of random noise in dA should be lower and also systematic errors from the calibration links are avoided.

More details of the excess phase calibration are given in several publications, for example, [38.59, 61]. Figure 38.12 shows the atmospheric phase delay and the corresponding amplitude (signal-to-noise ratio, SNR) for a typical TerraSAR-X occultation measurement. Typically, occultation measurements for the neutral atmosphere (0–120 km altitude) last around 1–2 min, and the atmospheric excess phase is around 1 km in the vicinity of the Earth's surface.

Derivation of Vertical Atmospheric Profiles

The calibration of the atmospheric excess phase can be regarded as a geodetic task and is the basis for the

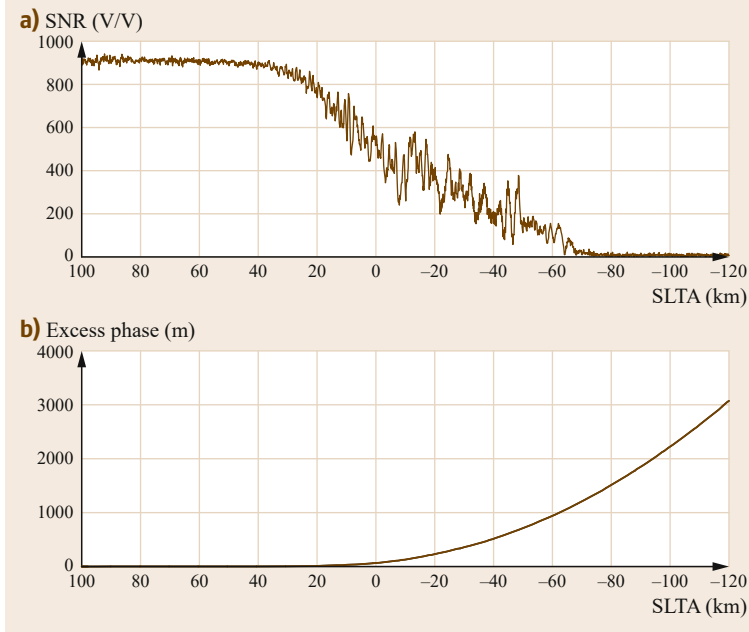


Fig. 38.12a,b Variation of the SNR (a) and the corresponding atmospheric excess phase (b) for a typical occultation event. SLTA indicates the Straight Line Tangent point Altitude. The plots characterize an occultation measurement of the TerraSAR-X satellite on February 18, 2012, 05:46 UTC, 79.6°N and 88.5°W (courtesy of F. Zus, GFZ)

mathematic-physical calculations to retrieve the vertical atmospheric profiles. The first step consists in the derivation of atmospheric bending angles, which are obtained from the time derivative of the atmospheric excess phase using the Doppler shift equation (e.g., [38.63])

$$f_d = f_c \left(\frac{c - (\mathbf{v}_2 \cdot \mathbf{m}_2)n_2}{c - (\mathbf{v}_1 \cdot \mathbf{m}_1)n_1} - 1 \right). \quad (38.15)$$

In (38.15), $\mathbf{v}_{1,2}$ are the velocity vectors of GPS and LEO satellite respectively, $\mathbf{m}_{1,2}$ are the unit wave vectors and $n_{1,2}$ are the refractivity at the satellite positions as shown in Fig. 38.13. The Doppler shift f_d is related to the phase L by

$$f_d = -\frac{f_c}{c} \frac{dL}{dt} \quad (38.16)$$

with the carrier frequency f_c and the vacuum light velocity c . L can be expressed as

$$L = L_0 + dA_{LO} \quad (38.17)$$

Thus, the Doppler shift f_d is represented by two terms

$$f_d = f_{d0} + f_{dA} \quad (38.18)$$

The first term in (38.18), f_{d0} , is equal to the frequency shift in the absence of the atmosphere and depends on L_0 . It can be calculated using the precise orbit information of the satellites (position and velocity). The

second term, f_{dA} , depends on the time derivative of the measured atmospheric excess phase A_{LO} of the occultation link between GPS and LEO. The bending angle is (Fig. 38.13)

$$\alpha = \phi_1 + \phi_2 + \theta - \pi. \quad (38.19)$$

ϕ_1 and ϕ_2 are unknowns; thus, one more equation is needed to calculate both ϕ_1 and ϕ_2 . Assuming local spherical symmetry of the refractivity $n = n(r)$, Snell's law applies

$$r_1 n(r_1) \sin \phi_1 = r_2 n(r_2) \sin \phi_2. \quad (38.20)$$

Equations (38.16) and (38.20) are nonlinear and cannot be solved analytically. It can be solved with an iterative method, as, for example, described in [38.63]. Starting with some increment $\Delta\phi_2 = \phi_2 - \phi_{20}$ (ϕ_{20} is equal to ϕ_2 in the absence of the atmosphere and can be calculated with satellite's orbit information) with (38.20) the corresponding $\Delta\phi_1$ can be calculated. Then the vectors \mathbf{m}_1 and \mathbf{m}_2 are constructed. Applying (38.15) $\Delta f = f_d - f_{d0}$ is calculated and compared with the observed value f_{dA} . Depending on the deviation of Δf from f_{dA} , the increment $\Delta\phi_2$ is modified and the procedure is repeated until ϕ_1 and ϕ_2 for each sample and with (38.20) the appropriate α is found.

The ionospheric correction is performed by the linear combination of the bending angle profiles obtained for each individual signal frequency (e.g., GPS L1 and

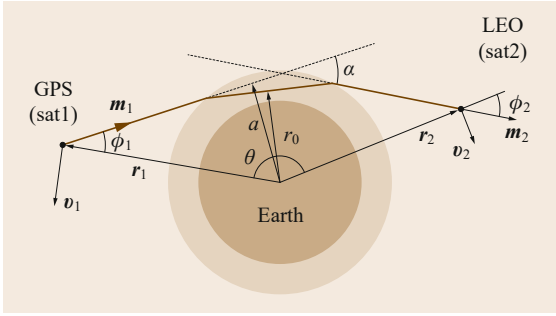


Fig. 38.13 Derivation of the bending angle α from the Doppler shift (a : impact parameter; r_0 : radius of the point of closest approach). For details, see the text

L2; [38.64]).

$$L_C(t) = \frac{f_1^2}{f_1^2 - f_2^2} L_1(t) - \frac{f_2^2}{f_1^2 - f_2^2} L_2(t) \quad (38.21)$$

$$\alpha_C(a) = \frac{f_1^2}{f_1^2 - f_2^2} \alpha_1(a) - \frac{f_2^2}{f_1^2 - f_2^2} \alpha_2(a). \quad (38.22)$$

The ionosphere correction in (38.22) avoids the effect of dispersion (L_1 and L_2 have separate signal paths), which forms the major error budget of (38.22) because the linear combination of the bending angles (38.22) is formed at the identical impact parameter for both frequencies.

Vertical profiles of the atmospheric refraction index n can then be retrieved from the ionosphere corrected bending angle profiles by the inverse Abel transform

$$n(r_0) = \exp \left(\frac{1}{\pi} \int_a^\infty \frac{\alpha(x)}{\sqrt{x^2 - a^2}} dx \right) \quad (38.23)$$

for the given point of the closest approach of the signal path to the Earth's surface r_0 , bending angle α , and impact parameter a .

After accounting for ionospheric bending as described above, the atmospheric refractivity ($N = (n-1) \cdot 10^6$) is related to pressure (p in mbar), temperature (T in K), and water vapor pressure (p_w in mbar) via the *Smith–Weintraub* equation [38.65]

$$N = 77.6 \frac{p}{T} + 3.73 \cdot 10^5 \frac{p_w}{T^2}. \quad (38.24)$$

For dry air, the density profiles are obtained from the known relationship between density and refractivity. Pressure and dry temperature

$$T_d = 77.6 \frac{p}{N} \quad (38.25)$$

are obtained from the hydrostatic equation and the equation of state for an ideal gas. There are numerous publications, describing these retrieval steps in very detail, for example, [38.54, 60, 66].

When water vapor is present, additional information is required to determine the humidity and density from refractivity profiles, due to the joint contribution of the dry and wet term to the refractivity in (38.24). Temperature profiles from operational meteorological analyses (e.g., of the European Centre for Medium-Range Weather Forecasts, ECMWF) are used to derive humidity profiles from the calculated refractivity in an iterative procedure [38.67]. This algorithm suffers from a high sensitivity to even small errors in the analyses temperatures, resulting in large uncertainties of the derived water vapor profiles [38.68]. More elaborate retrieval methods are based on the estimation of both temperature and humidity in parallel including the error characteristics of the measurement and the *background* information, which is usually obtained from meteorological analyses (optimal estimation, for example, [38.69, 70]). These methods show an increased potential for obtaining water vapor profiles with high accuracy.

By way of example, Fig. 38.14 shows vertical profiles of dry temperature and water vapor derived from a TerraSAR-X occultation measurement. The deviation (cold bias) of the dry temperature from the temperature below 10 km altitude is clearly seen and most obvious below 3 km, where the major part of the atmospheric water vapor is present. The magnitude of this deviation can itself be regarded as a measure for the atmospheric water vapor. It is noted that the key observables for the assimilation of RO data into forecast models are the bending angles or refractivities, rather than temperature and water vapor. The separation of these observables into dry and wet contributions, which finally provides the temperature and water vapor, is performed during the model analysis process using additional data from other observing systems. Also, for several climate change-related investigations, bending angle and refractivity data are used (e.g., [38.71]).

A major challenge in the GNSS RO data analysis is the parameter retrieval in the lower troposphere. Sharp refractivity gradients, mainly due to irregular water vapor distribution, complicate the proper signal tracking and the assumption of geometrical optics for analysis cannot be applied in contrast to higher altitudes. The application of the open-loop GNSS signal tracking technique as well as the development and application of advanced occultation data analysis techniques, however, brought significant progress during the last decade [38.72–76]. Another recent challenge is the data analysis in the upper stratosphere, where the occulta-

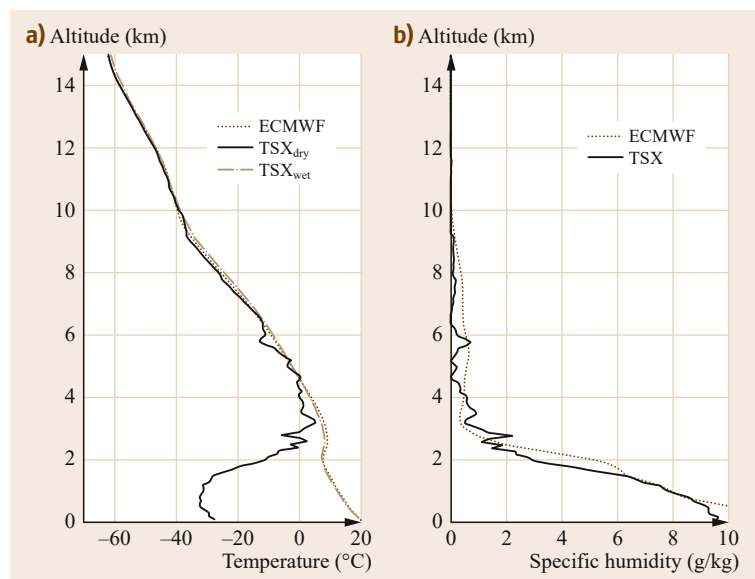


Fig. 38.14a,b Typical vertical dry (black) and wet (light brown) temperature (a) and water vapor profiles (b) derived from GNSS RO data. The example is from the TerraSAR-X (TSX) mission (July 19, 2010, 02:47 UTC, 31.43°N 145.67°E). Only the troposphere is shown. The RO data are compared with corresponding values from ECMWF analyzes (brown dotted line) (courtesy of S. Heise, GFZ)

tion signal is very weak and measurement errors (e.g., ionosphere) start to dominate the neutral atmosphere signal [38.66, 77].

38.2.3 Occultation Missions

A nearly complete and recent (as of 2016) list of satellite missions with GNSS RO instruments is given in [38.78]. Here, we give more details on selected missions of most importance for the development of the GNSS RO technique.

Initial GNSS RO data were recorded within the GPS/MET experiment aboard the MicroLab-1 satellite from 1995 to 1997 [38.52, 53]. However, the analysis of these data was primarily focused on the four *prime-times*, that is, periods of 2–3 weeks, when an anti-spoofing (A/S) encryption of the GPS signals was disabled and MicroLab-1 was oriented so that GPS satellites were occulted in the aft or anti-velocity direction toward the Earth's limb.

The German CHAMP satellite, launched on July 15, 2000 provided for the first time continuous and also near-real-time GPS RO data [38.57, 79]. These were especially used for various assimilation studies to investigate the potential improvement of RO data to numerical weather forecasts [38.80, 81]. In addition, CHAMP provided the first long-term set of GPS RO data covering the 2001–2009 period. In view of its high precision, it was used for initial climate-change-related investigations [38.82–84]. Furthermore, it triggered an international comparison of analysis results from different RO processing centers to define the structural uncertainty of GNSS RO data [38.85–87]. The CHAMP

RO experiment can therefore be regarded as a big success and as a forerunner for several succeeding missions. However, the daily number of available occultation measurements was limited to about 150.

Gravity Recovery And Climate Experiment (GRACE) is a US/German twin-satellite mission with focus to the detection of climate-relevant long-term variations of the Earth's gravity field determination, which was launched on March 17, 2002. The two spacecraft are equipped with the same *BlackJack* GPS RO flight receiver provided by Jet Propulsion Laboratory (JPL) as CHAMP. Continuous GPS RO measurements were activated on May 22, 2006 aboard the GRACE-A satellite [38.79], which provides around 130 near-real-time occultation profiles until today (as of end 2015). Recently, the GRACE Follow On (GRACE-FO) mission was confirmed, which is foreseen for launch in 2017 and will also include a GNSS-RO instrument according to the current planning.

A breakthrough for the number of daily occultations and for improved data quality in the lower troposphere was the launches of the U.S./Taiwan six-satellite-constellation FormoSAT-3/COSMIC (April 15, 2006; [38.55]) and of the two European Metop satellites (October 19, 2006 [38.88] and September 17, 2012). FormoSAT-3/COSMIC initially provided more than 2000 occultations daily in an open-loop tracking mode for better data quality in the lower troposphere. By 2013, the number of daily RO measurements had dropped to roughly half that value, since the nominal life time of the mission was reached and several satellites exhibit technical problems. Both Metop satellites together provide continuously and with high reliabil-

ity more than 1200 daily occultations. In addition, the German twin-satellite constellation TerraSAR-X (launched June 15, 2007) and TanDEM-X (launch June 21, 2010) provides a unique set of parallel occultation measurements [38.89] to investigate the accuracy potential of the GNSS RO technique in more detail. Data from TerraSAR-X are also provided in near real time for operational use in numerical weather forecasts.

The successor of the FormoSAT-3/COSMIC mission FormoSAT-7/COSMIC-2 is foreseen to be launched in 2016. This 12-satellite constellation will provide multi-GNSS LEO data (GPS, GLONASS, Galileo) from two different orbital inclinations. Six satellites are planned to be launched into low-inclination orbits in early 2016, and another six satellites into high-inclination orbits in 2018. This configuration will improve the global coverage of the GNSS RO data, especially in the Tropics. The GNSS RO payload, named TGRS for TriG (Tri-GNSS) GNSS RO System, is being developed by NASA's JPL and will be capable of tracking up to 12 000 high-quality profiles per day once both constellations are fully deployed. The third satellite of the Metop series will be launched in 2018. The planning for the follow-on system of the current EUMETSAT POLAR SYSTEM (EPS) includes also considerations for GNSS RO measurements. The EPS can be expected in the 2020 time frame.

In addition to these large and operational missions, there are several smaller international missions, which are in more detail overviewed in documents, generated by the International Radio Occultation Working Group (IROWG, [38.90]). The IROWG was established as a permanent Working Group of the Coordination Group for Meteorological Satellites (CGMS) in 2009 as part of the activities of the World Meteorological Organization

(WMO). The IROWG serves as a forum for operational and research users of RO data.

Another RO-related activity is CICERO (Community Initiative for Continuous Earth Remote Observation), which acts as a commercial provider of RO data. CICERO plans to launch a demonstration satellite in 2016 followed by an operational six-satellite constellation (CICERO-I) in the same year. Each satellite is foreseen to provide more than 900 GPS occultations per day.

CICERO-2, the planned extension up to 24 satellites by 2019, will offer enhanced performance with GPS/GLONASS/Galileo-enabled receivers. It potentially will provide more than 1600 occultations per day from each satellite.

38.2.4 Occultation Number and Global Distribution

Figure 38.15 shows the number of daily occultations from the six-satellite FormoSAT-3/COSMIC and Metop-A/B missions since the beginning of 2009. The maximum number of daily FormoSAT-3/COSMIC measurements was reached in early 2009 with up to 2500 profiles. At that time, the mission had already accomplished the nominal lifetime of three years. Nevertheless, it provided up to 2000 profiles daily even seven years after launch. The decreasing number of daily profiles and its quite large variation is associated with increasing technical problems of the satellites, which are already in orbit more than double of the nominal lifetime. The number of RO measurements, provided by both Metop satellites, is very stable, and around 1400 profiles daily are operationally available in near-real time.

The number of daily vertical profiles available from GRAS is before quality control, while the FormoSAT-

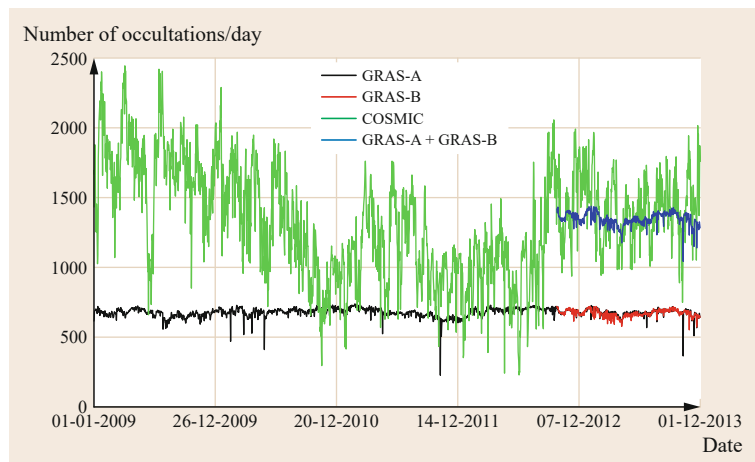


Fig. 38.15 Daily number of GNSS occultation measurements from GRAS-A (black), GRAS-B (red) and FormoSAT-3/COSMIC (green) between January 1, 2009 and December 1, 2013. The sum of the GRAS-A and -B data is indicated by the dark blue line (courtesy of A. von Engeln, EUMETSAT)

3/COSMIC numbers are after quality control. Typically, about 5–10% of the GRAS data are removed in quality control in the assimilation process of the weather services. Each of the GRAS receivers provides an almost constant number of daily occultations of around about 650–700. Some longer term variations are driven by the availability of GPS satellites for occultations, the short spikes are caused by, for example, loss of satellite data downloads or instrument updates.

Currently (as of end 2016), also GRACE-A and TerraSAR-X provide near-real-time RO data, but the daily number of around 150 measurements per satellite is much lower compared to Metop and FormoSAT-3/COSMIC.

A key property of the GNSS RO technique is the global coverage of the measurements, but the distribution is not uniform and depends mainly on the orbital geometry of the GNSS and the LEO satellites. As an example, Fig. 38.16 shows the global distribution of RO measurements from the FormoSAT-3/COSMIC mission (LEO orbit inclination $\approx 70^\circ$). The figure is based on about 4.2 millions RO profiles obtained between 2007 and 2012. The spatial distribution is global and nearly symmetric with respect to the Equator but not equally distributed. Most significant are the variations with latitude from ≈ 800 occultations per pixel (Equator) or even lower in the Polar regions to ≈ 2500 at 25 and 50° N/S in the mid-latitudes.

The orbit inclination of the LEO satellite is a key parameter to modify the global distribution of the RO measurements. Low-inclination LEO orbits will result in higher occultation density in the Equator region, which is of major interest for the prediction of severe weather events such as typhoons. Therefore, the first

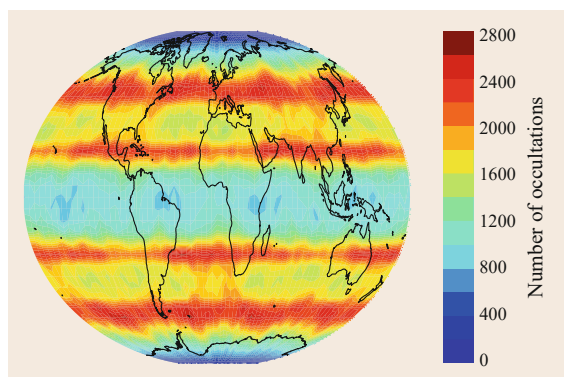


Fig. 38.16 Global distribution of GNSS RO data from the FormoSAT-3/COSMIC mission. The plot is based on about 4.2 million measurements, obtained between 2007 and 2012. The colors indicate the number of occultations per $5^\circ \times 5^\circ$ lat/lon grid cell (courtesy of C. Arras, GFZ)

six satellites of the 12 satellite FormoSAT-7/COSMIC-2 constellation (Sect. 38.2.3) will be deployed in a 20° inclination orbit. This will allow a higher equatorial occultation density compared to its predecessor mission. Near-polar orbiting LEO satellites (e.g., CHAMP, GRACE, Metop) exhibit a similar occultation distribution as shown in Fig. 38.16, but with more data available in the Polar regions (not shown here).

38.2.5 Measurement Accuracy

Numerous validation studies were performed throughout the last years to evaluate the quality of the various occultation missions (e.g., [38.52, 55, 79, 88, 89, 91]). The vertical profiles of refractivity, temperature, and water vapor were validated with data from different meteorological analyzes and radiosondes. In addition, co-located RO profiles, observed from two different satellite platforms, were compared [38.89, 92]. The results indicate that especially temperatures in the upper troposphere lower stratosphere (UTLS) region agree well with the analyses and sonde data.

Between approximately 8 and 25 km altitude, that is, in the UTLS region, mean temperature deviations are ≤ 1 K, and rms errors fall within the 1–2 K range. Also, only very small biases of about $\pm 0.1\%$ and rms uncertainties of $\leq 0.5\%$ are observed for the refractivity (see, for example, Fig. 38.17 for TerraSAR-X). The deviations at these heights could be either due to analysis/sonde data or the RO retrievals.

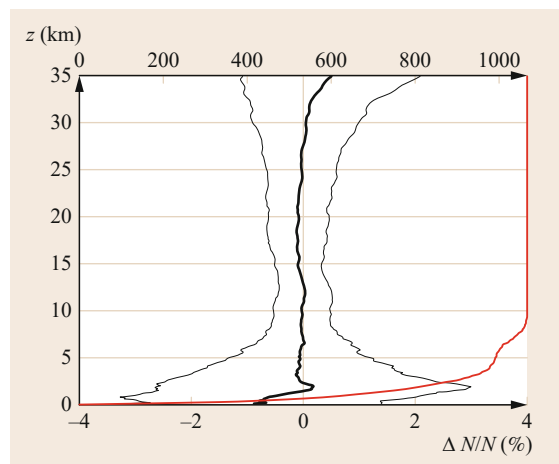


Fig. 38.17 Statistical comparison of refractivity profiles from TerraSAR-X with corresponding ECMWF data between November 26 and December 2, 2011. *Thick and thin black line* indicate bias and rms ($1 - \sigma$), the *red line* shows the number of compared data versus altitude (courtesy of F. Zus, GFZ)

A negative refractivity bias and significant loss of observations in the lower troposphere, especially at low latitudes, are observed in the RO retrievals and were in focus of numerous scientific studies, for example, [38.72–76]. One reason for this was found to be the application of the so-called phase lock loop (PLL) tracking mode of the occultation receivers of the early RO missions, such as GPS/MET or CHAMP. In the PLL mode, the phase of the RO signals is modeled (projected ahead) by extrapolating the previously extracted phase. This technique works well for standard GNSS observations (single-tone signals with sufficient SNR), but often fails for the occultation-geometry-like observations in the moist lower troposphere. Here, multipath propagation causes strong phase and amplitude fluctuations, which results in significant errors of the extrapolation-based phase model, a loss of SNR and finally loss of lock of the occultation signal. For this reason, the PLL mode does not allow in many cases a penetration of the occultation signals deep into the troposphere and is also the reason for systematic tracking errors. PLL mode receivers are also limited to the tracking of setting occultations only. An alternative tracking technique, open-loop (OL), that is, the raw sampling of the complex signal, was already applied to the data analysis of the planetary occultations and that also allows us to analyze rising occultation events [38.93]. However, a raw sampling of the signal is practically not feasible for routine GNSS RO sounding. Therefore, a model-based OL-tracking technique (e.g., [38.73]) was developed for the application in the moist troposphere for both rising and setting occultations and is used for several recent missions, for example, FormoSAT-3/COSMIC. In addition to these improvements of the occultation signal tracking, also the wave-optics-based retrieval techniques for the data analysis in the lower troposphere were improved during the years (e.g., [38.74–76, 94]).

The increased deviations above ≈ 25 km are also in the focus of recent investigations by RO and analysis specialists (e.g., [38.77]). On the one hand, GNSS RO retrievals become more difficult at these altitudes due to very small atmospheric excess phases; on the other hand, the analyses and the radiosondes exhibit problems at these altitudes.

The TerraSAR-X and TanDEM-X tandem satellite configuration (mean distance ≈ 20 km) provided for more than one year continuously vertical profiles at close quarters, recorded from different satellite platforms. This is a unique data set to investigate the accuracy potential of the RO technique and to determine its precision. Figure 38.18 shows a statistical comparison of corresponding refractivity profiles [38.89].

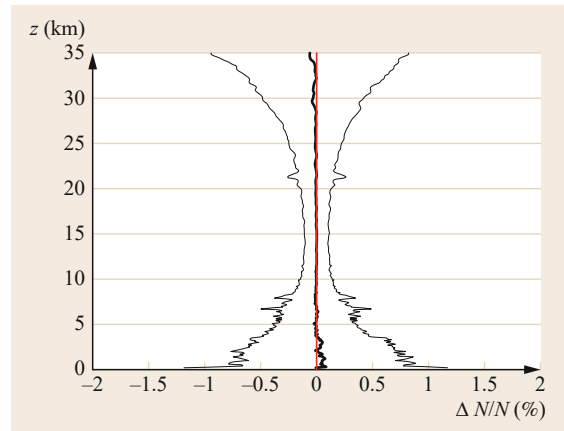


Fig. 38.18 Statistical comparison of the corresponding refractivity profiles from TerraSAR-X and TanDEM-X between November 26 and December 2, 2011. Thick and thin black lines indicate bias and rms, the red line indicates no deviation (courtesy of F. Zus, GFZ)

Nearly no or only an insignificant bias at the lower troposphere and above ≈ 30 km can be observed. The standard deviation is $\approx 0.1\%$ in the UTLS, $\approx 0.5\%$ in the lower troposphere, and above ≈ 30 km. These findings are in very good agreement with those in [38.92] for co-located FormoSAT-3/COSMIC profiles during the deployment phase of this multisatellite mission.

It is only briefly noted here that the high precision of the GNSS RO data is a valuable property to calibrate other satellite data from microwave sensors, which are widely used for global weather forecasts [38.95].

38.2.6 Prospects of New Navigation Satellite Systems

Similar to other GNSS applications, the availability of new navigation satellite systems (such as Galileo, BeiDou, and QZSS), along with the impressive renaissance of GLONASS and the modernization of GPS, will also be of great benefit for GNSS RO science and technology. GNSS RO will obviously profit from the significantly increased number of transmitting satellites. Even a single-satellite mission could potentially increase the number of daily occultation observations by a factor of 3 or 4, compared to only GPS. Besides this quantitative aspect, the new GNSS signal structures exhibit various advantages for the data quality of future RO missions. One example is the use of a third carrier frequency for improved ionospheric correction and better RO data quality in the stratosphere (Sect. 38.2.5). An initial summary on the prospects of the new GNSS for RO is given in [38.96].

38.2.7 Weather Prediction

The highlight of the GNSS RO applications and the breakthrough for its acceptance as an established atmospheric remote-sensing technique was the start of the operational use to improve global weather forecasts. Forerunner for this development was the German CHAMP satellite, which has been providing continuous near-real-time GNSS RO data since 2003. The average delay between the measurement and corresponding provision of globally distributed vertical atmospheric profiles was reduced from 5 h in 2003 to around 2 h in 2006 mainly by the implementation of optimized precise orbit determination procedures for CHAMP. These near-real-time data from GFZ were used by the leading forecast centers to develop appropriate assimilation techniques and to investigate and quantify the impact of the RO data on the forecasts [38.81]. Currently, the RO data are routinely used by the world-leading weather centers to improve their global numerical forecasts.

Several NWP (Numerical Weather Prediction) centers have reported a positive forecast impact with GNSS RO data (e.g., [38.80, 97–99]), despite the fact that the RO data numbers are low when compared with those of satellite radiances (major part of satellite data used) that are assimilated. For example, ECMWF assimilates around 10 million of conventional and satellite observations per 12 h period, of which 90% are satellite

radiance measurements, and only around 2% are GNSS RO-bending angles. The main GNSS RO impact is seen for upper-tropospheric and stratospheric temperatures. The GNSS RO measurements are beneficial because they provide complementary information to the satellite radiance measurements. Compared with satellite nadir sounders, the GNSS RO measurements have excellent vertical resolution and do not require bias correction, so they *anchor* the bias correction applied to satellite radiances and help identify NWP model biases [38.100].

Figure 38.19 shows the historical begin of the assimilation of GNSS RO data from FormoSAT-3/COSMIC, CHAMP, and GRACE-A on December 12, 2006. The major information is the reduction/elimination of the ECMWF bias in the background and analysis temperature (≈ 0.2 K and ≈ 0.4 K) and geopotential height (5–10 m) of the 100 hPa pressure level compared to radiosonde data, which can be regarded as truth at these altitudes.

38.2.8 Climate Monitoring

RO observations are well suited for establishing a stable, long-term record required for climate monitoring (e.g., [38.82, 101, 102]). Key properties for this application are: global coverage, high accuracy, high vertical resolution, and independence from weather. Most important, however, is that the fundamental RO obser-

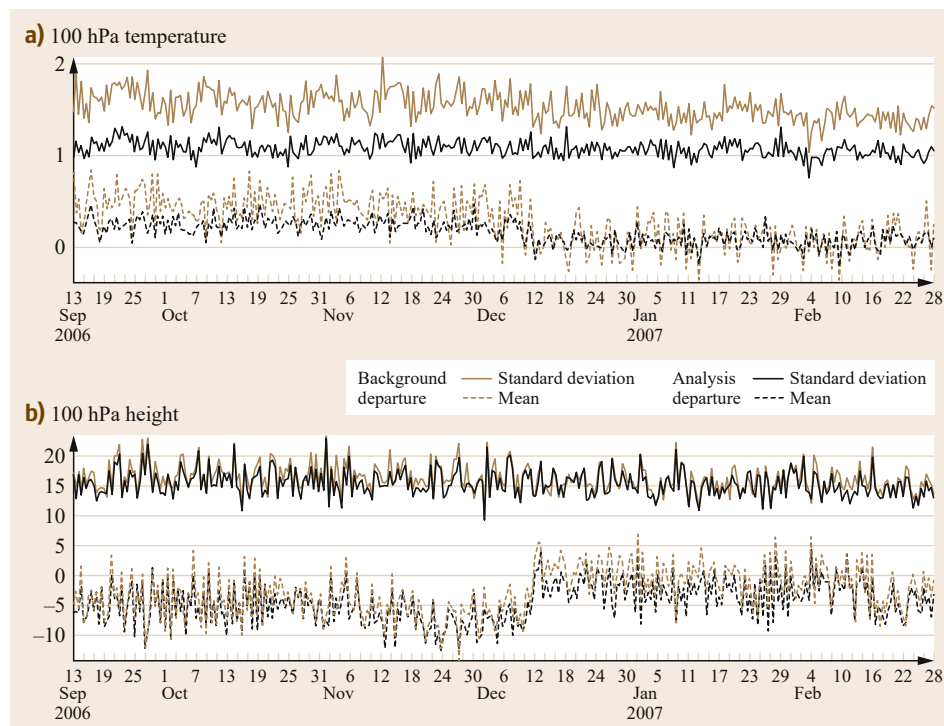


Fig. 38.19a,b Time series of the mean and standard deviation of the ECMWF operational background departures and analysis departures for (a) temperature and (b) geopotential height radiosonde measurements at 100 hPa in the southern hemisphere. GNSS-RO was introduced on December 12, 2006 (courtesy of S. Healy, ECMWF)

vation is a measurement of time (determination of the signal travel time), which is performed by atomic clocks with unequaled accuracy and stability. During an occultation, the GNSS receiver measures the change in the flight time of the signal transmitted by the occulted GNSS satellite. The clocks aboard the GNSS transmitters remain synchronized to the most stable atomic clocks on the ground. The clock in a GNSS receivers aboard an LEO satellite is synchronized, using the signals from up to 10 nonocculted GNSS transmitters in view and is thus tied to the stable ground-based GNSS time as well. Therefore, an extremely accurate measurement of the signal flight time with long-term stability can be achieved. Because the fundamental observation is a measurement of time, RO is a promising technique for climate monitoring.

The detection of climate trends is enormously important, especially because of their huge social and economical consequences, but there is presently no atmospheric instrument that can meet the stringent climate monitoring requirements of 0.5 K accuracy and 0.04 K decade⁻¹ stability [38.102].

Global Temperature Trends

The upper troposphere and lower stratosphere (UTLS) are the key regions of the atmosphere with important links to the stratosphere–troposphere exchange as well as climate research. The determination of the UTLS temperature and tropopause (TP) height trends is crucial for the monitoring of climate-change processes (Sect. 38.2.8). Global high-resolution temperature observations in the UTLS region are only available from GNSS RO data. Here the CHAMP mission has generated the first long-term RO data set (2001–2008) that is continued with data from other mission (GRACE, FormoSAT-3/COSMIC, MetOp, TerraSAR-X). The UTLS region is also the vertical atmosphere region, where GNSS RO exhibits the highest accuracy (Sect. 38.2.5), another important property for the use in UTLS climate studies.

A global pattern of atmospheric temperature trends between 5 and 25 km is shown in Fig. 38.20. The figure was derived from CHAMP, GRACE-A, and TerraSAR-X GPS RO data between 2001 and 2013. The TP altitude is indicated with a white line. A slight overall warming in the upper troposphere (above 5 km to the TP) can be observed with largest values in the subtropical region of the southern hemisphere (SH). In the lower stratosphere from the TP up to 25 km predominant negative temperature trends (cooling) are detected. The equatorial TP region and the lower SH stratosphere reveal warming [38.82, 83, 103].

The results of these studies indicate the great potential of the very precise GNSS data to monitor even

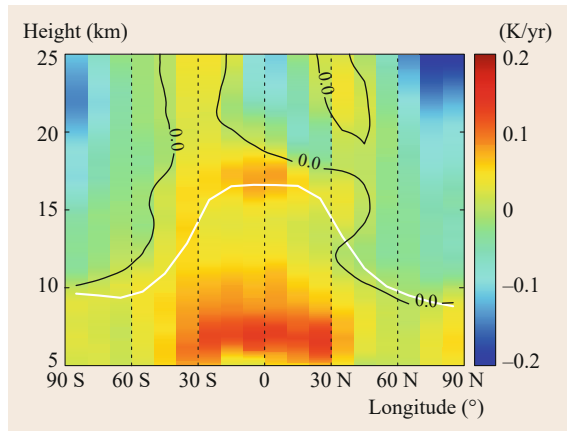


Fig. 38.20 Global temperature trends in the upper troposphere and lower stratosphere based on CHAMP, GRACE, and TerraSAR-X GPS RO data (2001–2013). The *solid white line* denotes the mean TP height (courtesy of T. Schmidt, GFZ)

small atmospheric temperature trends. This is also one reason for the current use of GPS RO data to validate a new model system for of mid-term climate forecast, which is, for example, currently developed in Germany.

To ensure a high data quality, especially for the RO climate applications, the international RO science community started in 2009 an important activity. RO products from different processing centers are compared regularly for the determination of the structural uncertainty in climate data records and the stability of trends (Fig. 38.21). These multicenter-based results ensure a high quality of the RO data analysis and provide more complete and reliable climatological information as derived from the results of only one center [38.85, 86].

The Tropopause: Indicator for Climate Change

The TP region separates the troposphere and stratosphere that have fundamentally different characteristics with respect to chemical composition and static stability. Therefore, the determination of TP parameters, such as altitude or temperature, on a global scale is an important goal for many branches in atmospheric research [38.104]. With regard to the current climate change discussion, TP parameters have received more attention in recent years since they are used to describe climate variability and change. The global mean TP altitude shows an increase in re-analyses and radiosonde observations over the last decades and seems to be a more sensitive indicator for climate change than the Earth's surface temperature [38.105]. Another ap-

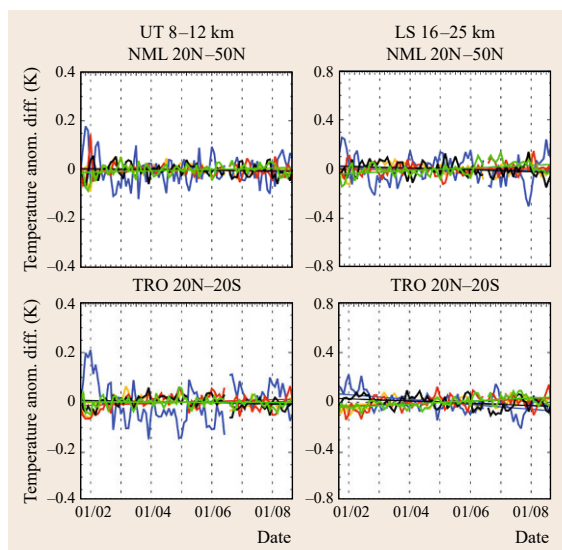


Fig. 38.21 Structural uncertainty in RO temperature records from different processing centers: DMI Copenhagen (yellow), GFZ Potsdam (blue), JPL Pasadena (red), UCAR Boulder (black), and WEGC Graz (green). Shown are difference time series of temperature anomalies for each center (with respect to the all-center mean) for the upper troposphere (left) and the lower stratosphere (right), for northern mid-latitudes (top) and the tropics (bottom). The overplotted difference trends are close to zero and indicate the stability of the RO data record (courtesy of A. Steiner, Wegener Center)

plication area of TP studies deals with the role of the TP region in the stratosphere–troposphere exchange. In this context, multiple tropopauses or TP break regions are important because in these regions most of the exchange processes take place (e.g., [38.106]).

One important data source for the determination of TP parameters is radiosondes. Despite good vertical resolution of the radiosonde data, global coverage is impossible. In contrast, the RO technique offers both global coverage and good vertical resolution as well and is therefore of particular relevance for detailed TP studies. First GNSS RO results for the tropical TP region were already published based on GPS/MET data [38.107, 108]. An example for the TP-related investigations is shown in Fig. 38.22 based on the investigations described in [38.109]. A significant increase of the global mean TP height of about 6 m/year between 2001 and 2011 was found associated with a warming in the upper troposphere. This could be an indication for a warming (extension) of the entire TP (connected with the cooling of the stratosphere), but longer data sets are needed to get more confidence of these early GNSS RO results for climate change research.

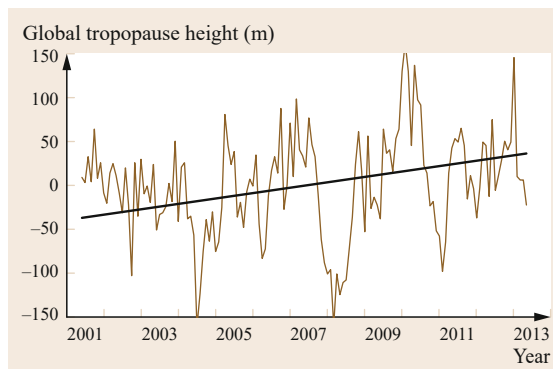


Fig. 38.22 Global TP height trend (black line) based on CHAMP, GRACE, and TerraSAR-X GPS RO data (2001–2011). The brown line indicates the monthly mean global TP height, derived from these RO missions (courtesy of T. Schmidt, GFZ)

Gravity Waves

Another important and climate change detection related application of the GNSS RO data is the derivation of atmospheric wave parameters on a global scale. Most relevant in this respect are gravity waves (GWs, wave phenomena, where the force of gravity tries to restore equilibrium), which play an important role for the general atmospheric circulation due to the related transport of energy and momentum between different regions of the atmosphere. Therefore, their analysis is of great interest for local weather forecasts and global climate modeling.

Early studies were initiated with GPS/MET data [38.110] and focused on vertically propagating waves. Recent studies (e.g., [38.111]) use much larger databases from multisatellite constellations and indicate the potential to derive also horizontal wave properties. Figure 38.23 shows the momentum flux (MF) distribution for July as a mean of 4 years (2007–2010) within the altitude range of 20–25 km. High values of MF along the southern Andes and at the east of the Andes are due to strong steady west wind crossing the mountains in that region. In the tropical regions, the high gravity wave activity, which induces high values of MF, is due to intense convection. The northern hemisphere is rather quiet in the local summer, therefore showing low MF values (Fig. 38.23). Atmospheric waves, for example, tides, can also be detected in spatiotemporal signatures of ionospheric irregularities (Sporadic E layer), detected using the GNSS RO technique [38.112].

The Planetary Boundary Layer

Recently scientists recognized more and more importance of the dynamics of the planetary boundary layer

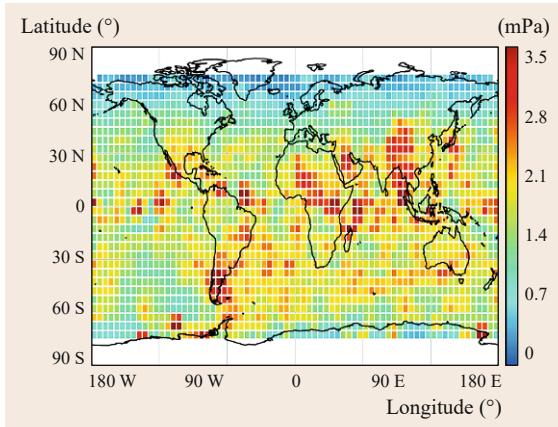


Fig. 38.23 Horizontal momentum flux (MF, a measure for horizontal energy transport by gravity waves) distribution generated from groupings of three co-located GPS RO profiles from the FormoSAT-3/COSMIC mission as 4-year mean values (2007–2010) for July for the altitude range of 20–25 km (courtesy of A. Faber, GFZ)

(PBL) inversions to the overall climate system. The PBL is the lowermost atmospheric layer directly affected by the Earth's surface. Commonly, the boundary between this turbulently mixed layer and the stably stratified atmosphere above is characterized by a temperature inversion and the decrease of relative and absolute humidity, especially in the moist tropics and subtropics. The top of the PBL is sharper and horizontally more homogeneous in the subtropics, where it is often called a trade wind inversion, than in the tropics and over oceans compared to land. The depth of the PBL is an important parameter for numerical weather prediction and climate models. The key property of GNSS RO

for monitoring the PBL is global coverage and high vertical resolution; especially, the top of the PBL is associated with sharp vertical refractivity gradients, which can be clearly identified with GNSS RO (Fig. 38.24).

Several studies from different research groups were performed in recent years [38.94, 113, 114]. For example, in [38.113], three-year climatologies of mean PBL heights, derived from GNSS RO (Fig. 38.24) and ECMWF Reanalysis Interim (ERA-Int), show similar spatial and seasonal variations, but the GNSS RO heights were higher by 500 m, and the standard deviation was also higher from GNSS RO, especially in the tropics, which was analyzed in more detail for various regions, as the Pacific Ocean and the Sahara desert. The results suggest that the underlying causes of the bias between GNSS RO and ERA-Int likely vary from region to region. Another main result of this study is the statement that GNSS RO profiles actually contain vertically resolved information above and within the PBL, information which can be difficult to obtain through any other satellite measurement.

38.2.9 Synergy of GNSS Radio Occultation with Reflectometry

Recently, GNSS signals reflected off the surface of the Earth are in focus of intense international GNSS research (GNSS-Reflectometry, GNSS-R). These signals promise a broad range and numerous geophysical applications for remote sensing (e.g., [38.115, 116]) and are in more detail focused in Chap. 40 of this book. Whereas GNSS atmosphere sounding is already fully recognized as an established atmospheric remote-sensing technique, GNSS-R needs more and concentrated international research to exploit its full and

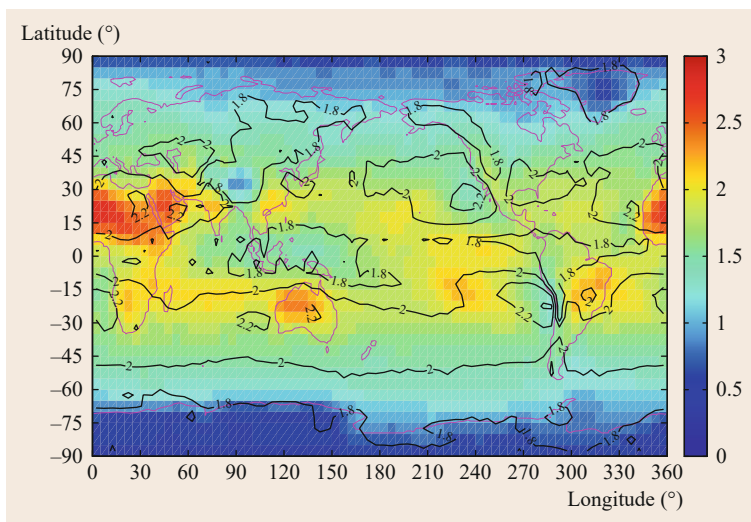


Fig. 38.24 Height of the mean global PBL (resolution $5 \times 5^\circ$, derived from five years of GPS RO data from the FormoSAT-3/COSMIC mission (2007–2011) (courtesy of C.O. Ao, JPL)

unique potential for the remote sensing of water, ice, and land surfaces but also for atmosphere/ionosphere sounding. Important milestones for this development are the dedicated satellite missions CYGNSS (CYclone Global Navigation Satellite System, National Aeronautics and Space Administration (NASA), [38.117]) and GEROSS (GNSS Reflectometry Radio Occultation and Scatterometry aboard the International Space Station; European Space Agency (ESA) [38.115]), which focus on a global application of the GNSS reflectometry.

Part of these activities is the investigation of the potential of the carrier-phase interferometry between the reflected and direct occultation signals (coherent reflectometry), which was initially demonstrated using the measurements from GPS/MET and CHAMP [38.118–120], see Fig. 38.25. These studies indicated that sub-meter sensitivity on the surface heights can formally be reached with this technique, which offers potential altimetric applications of ocean and ice surfaces. Improved and specific GNSS-tracking software probably can improve this reached accuracy in future with combined occultation/reflection experiments. A major advantage of the coherent reflectometry compared to nadir-viewing reflectometry is that only a low-gain limb-viewing antenna is required, which allows the application also aboard small satellites in future GNSS

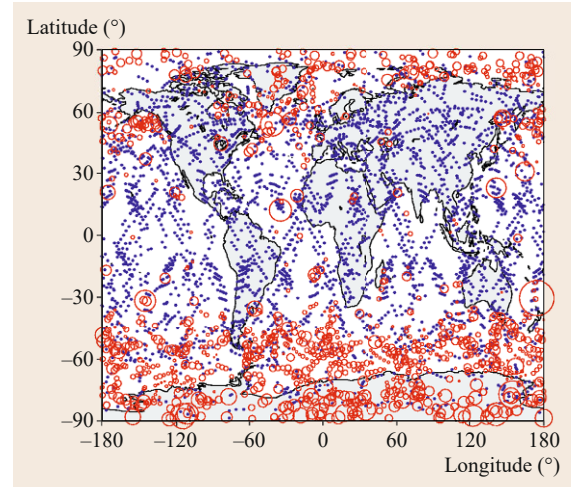


Fig. 38.25 Geographical distribution of 3783 occultation events observed between 14 May and 10 June 2001. *Blue dots* indicate 2571 observations without reflection signatures; 1212 reflection events are marked as *red circles*. Circle diameter is proportional to the reflected intensity (courtesy of G. Beyerle, GFZ)

remote-sensing constellations. But for such application, a more detailed evaluation of the accuracy potential of the coherent reflectometry is required.

38.3 Outlook

Ground- and satellite-based atmosphere sounding techniques with their broad spectrum of applications, especially in weather forecast and climate change-related research, were introduced. An overview of the main applications is presented in Table 38.2. Today many thousands of continuously operating ground-based stations exist. Some hundreds are coordinated globally by the IGS and many more are coordinated on a regional or local (national) level. For real-time applications such as weather forecasting, the most important information is the temporal variations, the timing of moving air masses. This does not necessarily require an absolute calibration and homogeneous processing. In this aspect, the monitoring of the water vapor content over decades, for climate change studies, is much more demanding in terms of homogeneous networks and processing in order to draw correct conclusions for the very tiny trends that are expected.

The ground-based data have so far mainly been acquired from sites on land. Platforms on ships or bouys, have more logistics involved and suffer of some extent from the need for simultaneous estimates of the vertical coordinate. The concept has been demonstrated

(see [38.121, 122]) and assuming more efficient data communication in the future this may offer a much improved global coverage.

Beginning with the initial and very successful measurements from GPS/MET and by the follow-on missions, as, for example, CHAMP, GRACE, FormoSAT-3/COSMIC and Metop, the innovative GNSS RO method became an established atmospheric remote-sensing technique within the last two decades. A series of new missions is planned and will be realized within the coming years. Our general conclusion is that GNSS atmosphere sounding, ground as well as satellite based, underwent a revolutionary development especially during the last decade and are now fully recognized atmospheric remote-sensing techniques. This development is documented by a broad variety of scientific and also operational applications, most visible is the continuous use of ground- and satellite-based GNSS data for the improvement of numerical weather forecasts since 2006.

The recent GNSS developments will further push these activities. There are upcoming and modernized transmitter systems, continuously increasing receiver

Table 38.2 GNSS meteorology applications for remote sensing of the neutral atmosphere

Application	Ground-based receivers	Satellite-based receivers
Weather forecasting		
IWV timeseries	Yes	–
Bending angle/Refractivity profiles	–	Yes
Climate		
IWV trends	Yes	–
Annual and diurnal IWV signals	Yes	–
Global Temperature trends	–	Yes
Tropopause characteristics	–	Yes
Atmospheric research		
Atmospheric convection regional scale	Yes	–
Atmospheric waves	–	Yes
Complement other sensors (e.g., infrared)	–	Yes
Large scale atmospheric circulation	–	Yes
Planetary boundary layer	–	Yes
Tropical cyclones	–	Yes
Upper troposphere and lower stratosphere	–	Yes
3D-Distributions of water vapor	Yes	–
Spatiotemporal variability in the IWV	Yes	–

infrastructure, with the extension of ground networks and more GNSS flight receivers, but probably also new marine, ground and flight platforms in the near future, not to forget the *everyones* receivers such as smartphones.

These developments will not only increase the number of atmospheric GNSS measurements, but probably also will allow a better data quality. It will further stimulate the existing applications but probably also will open the door for new and innovative applications. GNSS atmosphere sounding is a story of success, which will be continuously updated.

Acknowledgments. The authors want to thank several colleagues for providing information and figures: Chi On Ao (JPL), Christina Arras (GFZ), Georg Beyerle (GFZ), Galina Dick (GFZ), Axel von Engeln (EUMETSAT), Antonia Faber (GFZ), Rüdiger Haas (Chalmers), Sean Healy (ECMWF), Stefan Heise (GFZ), Tong Ning (Swedish Mapping, Cadastral and Land Registration Authority), Torsten Schmidt (GFZ), Hans-Georg Scherneck (Chalmers), Bill Schreiner (UCAR), Andrea Steiner (Wegener Center), Tom Yunck (GeoOptics), and Florian Zus (GFZ).

References

- 38.1 D.M. Tralli, T.H. Dixon, S.A. Stephens: Effect of wet tropospheric path delays on estimation of geodetic baselines in the Gulf of California using the Global Positioning System, *J. Geophys. Res.* **93**(B6), 6545–6557 (1988)
- 38.2 D.M. Tralli, S.M. Lichten: Stochastic estimation of tropospheric path delays in Global Positioning System geodetic measurements, *Bull. Geod.* **64**, 127–159 (1990)
- 38.3 Y.E. Bar-Sever, P.M. Kroger, A.J. Börjesson: Estimating horizontal gradients of tropospheric path delay with a single GPS receiver, *J. Geophys. Res.* **103**, 5019–5035 (1998)
- 38.4 P. Elsegui, J.L. Davis: Accuracy assessment of GPS slant-path determinations, *Int. Workshop GPS Meteorol.*, Tsukuba, ed. by T. Iwabuchi, Y. Shoji (Meteorological Society of Japan, Tsukuba 2004) pp. 1–6
- 38.5 A.E. Niell: Global mapping functions for the atmosphere delay at radio wavelengths, *J. Geophys. Res.* **101**(B2), 3227–3246 (1996)
- 38.6 I.D. Thomas, M.A. King, P.J. Clarke, N.T. Penna: Precipitable water vapor estimates from homogeneously reprocessed GPS data: An intertechnique comparison in Antarctica, *J. Geophys. Res.* **116**, D19101 (2011)
- 38.7 T.A. Herring: Precision of vertical position estimates from very long baseline interferometry, *J. Geophys. Res.* **91**, 9177–9182 (1986)
- 38.8 S. Kedar, G.A. Hajj, B.D. Wilson, M.B. Heflin: The effect of the second order GPS ionospheric correction on receiver positions, *Geophys. Res. Lett.* **30**(16), 1829 (2003)
- 38.9 M. Hernandez-Pajares, J.M. Juan, J. Sanz, R. Ors: Second-order ionospheric term in GPS: Implementation and impact on geodetic estimates, *J. Geophys. Res.* **112**, B08417 (2007)

- 38.10 E.J. Petrie, M.A. King, P. Moore, D.A. Lavallè: Higher-order ionospheric effects on the GPS reference frame and velocities, *J. Geophys. Res.* **115**(10), B03417 (2013)
- 38.11 R. Schmid, P. Steigenberger, G. Gendt, M. Ge, M. Rothacher: Generation of a consistent absolute phase center correction model for GPS receiver and satellite antennas, *J. Geod.* **81**(5), 781–798 (2007)
- 38.12 P.O. Jarlemark, T.R. Emardson, J.M. Johansson, G. Elgered: Ground-based GPS for validation of climate models: The impact of satellite antenna phase center variations, *IEEE Trans. Geosci. Remote Sens.* **GE-48**(10), 3847–3854 (2010)
- 38.13 T. Ning, G. Elgered, J.M. Johansson: The impact of microwave absorber and radome geometries on GNSS measurements of station coordinates and atmospheric water vapour, *Adv. Space Res.* **47**, 186–196 (2011)
- 38.14 National Institute of Standards and Technology, US, <http://physics.nist.gov/cgi-bin/cuu/Value?r>
- 38.15 J.L. Davis, T.A. Herring, I.I. Shapiro, A.E.E. Rogers, G. Elgered: Geodesy by radio interferometry: Effects of atmospheric modeling errors on estimates of baseline length, *Radio Sci.* **20**(6), 1593–1607 (1985)
- 38.16 S. Heise, G. Dick, G. Gendt, T. Schmidt, J. Wickert: Integrated water vapor from IGS ground-based GPS observations: Initial results from a 5-min data set, *Ann. Geophysicae* **27**, 2851–2859 (2009)
- 38.17 T. Nilsson, G. Elgered: Long-term trends in the atmospheric water vapor content estimated from ground-based GPS data, *J. Geophys. Res.* **113**(D19101), 1–12 (2008)
- 38.18 S. Vey, R. Dietrich, M. Fritsche, A. Rülke, P. Steigenberger, M. Rothacher: On the homogeneity and interpretation of precipitable water time series derived from global GPS observations, *J. Geophys. Res.* **114**(D10101), 1–15 (2009)
- 38.19 K. Lagler, M. Schindelegger, J. Böhm, H. Krásná, T. Nilsson: GPT2: Empirical slant delay model for radio space geodetic techniques, *Geophys. Res. Lett.* **40**, 1069–1073 (2013)
- 38.20 R.B. Stull: *An introduction +to boundary layer meteorology*, 3rd edn. (Academic Press, San Diego 1992)
- 38.21 J.P. Hauser: Effects of deviations from hydrostatic equilibrium on atmospheric corrections to satellite and lunar laser range measurements, *J. Geophys. Res.* **94**, 10182–10186 (1991)
- 38.22 M. Bevis, S. Chiswell, T.A. Herring, R.A. Anthes, C. Rocken, R.H. Ware: GPS meteorology: Mapping zenith wet delays onto precipitable water, *J. Appl. Meteorol.* **33**, 379–386 (1994)
- 38.23 J. Wang, L. Zhang, A. Dai: Global estimates of water-vapor-weighted mean temperature of the atmosphere for GPS applications, *J. Geophys. Res.* **110**(D21101), 1–17 (2005)
- 38.24 T.R. Emardson, G. Elgered, J.M. Johansson: Three Months of continuous monitoring of atmospheric water vapor with a network of Global Positioning System receivers, *J. Geophys. Res.* **103**, 1807–1820 (1998)
- 38.25 T.R. Emardson, H.J.P. Derks: On the Relation Between the Wet Delay and the Integrated Precipitable Water Vapour in the European Atmosphere, *Meteorol. Appl.* **7**, 61–68 (2000)
- 38.26 G.S. Kell: Density, thermal expansivity, and compressibility of liquid water from 0 to 150 °C: Correlations and tables for atmospheric pressure and saturation reviewed and expressed on 1968 temperature scale, *J. Chem. Engineering Data* **20**(1), 97–105 (1975)
- 38.27 G. Elgered, J.M. Johansson, B.O. Rönnäng, J.L. Davis: Measuring regional atmospheric water vapor using the Swedish permanent GPS network, *Geophys. Res. Lett.* **24**, 2663–2666 (1997)
- 38.28 J.L. Davis, G. Elgered: The spatio-temporal structure of GPS water-vapor determinations, *Phys. Chem. Earth* **23**(1), 91–96 (1998)
- 38.29 T.R. Emardson, F.H. Webb: Estimating the motion of atmospheric water vapor using the Global Positioning System, *GPS Solutions* **6**, 58–64 (2002)
- 38.30 EUMETNET, The Network of European Meteorological Services, <http://egvap.dmi.dk/>
- 38.31 H.-S. Bauer, V. Wulfmeyer, T. Schwitalla, F. Zus, M. Grzeschik: Operational assimilation of GPS slant path delay measurements into the MM5 4DVAR system, *Tellus* **63A**, 263–282 (2011)
- 38.32 P. Poli, P. Moll, F. Rabier, G. Desroziers, B. Chapnik, L. Berre, S.B. Healy, E. Andersson, F.-Z. El Gue-lai: Forecast impact studies of zenith total delay data from European near real-time GPS stations in Météo France 4DVAR, *J. Geophys. Res.* **112**(D06114), 1–16 (2007)
- 38.33 M. Bender, G. Dick, M. Ge, Z. Deng, J. Wickert, H.-G. Kahle, A. Raabe, G. Tetzlaff: Operational assimilation of GPS zenith total delay observations into the Met Office numerical weather prediction models, *Mon. Weather Rev.* **140**, 2706–2719 (2012)
- 38.34 R. Eresmaa, H. Järvinen: An observation operator for ground-based GPS slant delays, *Tellus* **58A**, 131–140 (2006)
- 38.35 T. Nilsson, L. Gradinarsky: Water vapor tomography using GPS phase observations: Simulation results, *IEEE Trans. Geosci. Remote Sens.* **GE-44**(10), 2927–2941 (2006)
- 38.36 M. Bender, G. Dick, M. Ge, Z. Deng, J. Wickert, H.-G. Kahle, A. Raabe, G. Tetzlaff: Development of a GNSS water vapour tomography system using algebraic reconstruction techniques, *Adv. Space Res.* **47/10**, 1704–1720 (2011)
- 38.37 A. Flores, G. Ruffini, A. Rius: 4D tropospheric tomography using GPS slant wet delays, *Annal. Geophysicae* **18**, 223–234 (2001)
- 38.38 S.A. Buehler, A. von Engel, E. Brocard, V.O. John, T. Kuhn, P. Eriksson: Recent developments in the line-by-line modeling of outgoing longwave radiation, *J. Quant. Spectrosc. Radiat. Transfer* **98**(3), 446–457 (2006)
- 38.39 N.D. Gordon, A.K. Jonko, P.M. Forster, K.M. Shell: An observationally based constraint on the water-vapor feedback, *J. Geophys. Res.* **118**, 12435–12443 (2014)

- 38.40 D.J. Seidel, F.H. Berger, H.J. Diamond, J. Dykema, D. Goodrich, F. Immler, W. Murray, T. Peterson, D. Sisterson, M. Sommer, P. Thorne, H. Vömel, J. Wang: Reference upper-air observations for climate: Rationale, progress, and plans, *Bull. Am. Meteorol. Soc.* **90**, 361–369 (2009)
- 38.41 H.-G. Scherneck, J.M. Johansson, H. Koivula, T. van Dam, J.L. Davis: Vertical crustal motion observed in the BIFROST project, *J. Geodyn.* **35**, 425–441 (2003)
- 38.42 T. Ning, G. Elgered: Trends in the atmospheric water vapor content from ground-based GPS: The impact of the elevation cutoff angle, *IEEE J-STARS* **5**(3), 744–751 (2012)
- 38.43 S. Jin, O.F. Luo: Variability and Climatology of PWV From Global 13-Year GPS Observations, *IEEE Trans. Geosci. Remote Sens.* **GE-47**, 1918–1924 (2009)
- 38.44 J.P. Ortiz de Galisteo, Y. Bennouna, C. Toledano, V. Cachorro, P. Romero, M.I. Andrés, B. Torres: Analysis of the annual cycle of the precipitable water vapour over Spain from 10-year homogenized series of GPS data, *Quart. J. Roy. Meteorol. Soc.* **139**, 948–958 (2013)
- 38.45 Y.S. Bennouna, B. Torres, V.E. Cachorro, J.P. Ortiz de Galisteo, C. Toledano: The evaluation of the integrated water vapour annual cycle over the Iberian Peninsula from EOS-MODIS against different ground-based techniques, *Quart. J. Roy. Meteorol. Soc.* **139**, 1935–1956 (2013)
- 38.46 S. Jin, O.F. Luo, S. Gleason: Characterization of diurnal cycles in ZTD from a decade of global GPS observations, *J. Geod.* **83**, 537–545 (2009)
- 38.47 J. Wang, L. Zhang: Climate applications of a global, 2-hourly atmospheric precipitable water dataset derived from IGS tropospheric products, *J. Geod.* **83**, 209–217 (2009)
- 38.48 S. Pramualsakkikul, R. Haas, G. Elgered, H.-G. Scherneck: Sensing of diurnal and semi-diurnal variability in the water vapour content in the tropics using GPS measurements, *Meteorol. Appl.* **14**, 403–412 (2007)
- 38.49 E. Jakobson, H. Övrlid, G. Elgered: Diurnal variability of precipitable water in the Baltic region, impact on transmittance of the direct solar radiation, *Boreal Environment Res.* **14**, 45–55 (2009)
- 38.50 S. Vey, R. Dietrich, A. Rülke, M. Fritsche, P. Steigenberger, M. Rothacher: Validation of precipitable water vapor within the NCEP/DOE reanalysis using global GPS observations from one decade, *J. Climate* **23**, 1675–1695 (2010)
- 38.51 T. Ning, G. Elgered, U. Willén, J.M. Johansson: Evaluation of the atmospheric water vapor content in a regional climate model using ground-based GPS measurements, *J. Geophys. Res.* **118**, 329–339 (2013)
- 38.52 C. Rocken, R. Anthes, M. Exner, D. Hunt, S. Sokolovskiy, R. Ware, M. Gorbunov, W. Schreiner, D. Feng, B. Herman, Y.-H. Kuo, X. Zou: Analysis and validation of GPS/MET data in the neutral atmosphere, *J. Geophys. Res.* **102**, 29849–29866 (1997)
- 38.53 R. Ware, D. Exner, M. Feng, M. Gorbunov, K. Hardy, B. Herman, Y. Kuo, T.K. Meehan, W.G. Melbourne, C. Rocken, W. Schreiner, S. Sokolovskiy, F. Solheim, X. Zou, R. Anthes, S. Businger, K. Trenberth: GPS sounding of the atmosphere from low Earth orbit – Preliminary results, *Bull. Am. Met. Soc.* **77**, 19–40 (1996)
- 38.54 E.R. Kursinski, G.A. Hajj, K.R. Hardy, J.T. Schofield, R. Linfield: Observing the Earth's atmosphere with radio occultation measurements using the Global Positioning System, *J. Geophys. Res.* **102**, 23429–23465 (1997)
- 38.55 R.A. Anthes, P.A. Bernhardt, Y. Chen, K. Cucurull, K.F. Dymond, S. Ector, S.B. Healy, S.-P. Ho, D.C. Hunt, Y.-H. Kuo, H. Liu, K. Manning, C. McCormick, T.K. Meehan, W.J. Randel, C. Rocken, W.S. Schreiner, S.V. Sokolovskiy, S. Syndergaard, D.C. Thompson, K.E. Trenberth, T.-K. Wee, N.L. Yen, Z. Zhang: The COSMIC/Formosat-3 Mission: Early results, *Bull. Am. Met. Soc.* **89**(3), 313–333 (2008)
- 38.56 S. Healy: Operational assimilation of GPS radio occultation measurements at ECMWF, *ECMWF Newsletter* 111 (2007)
- 38.57 J. Wickert, C. Reigber, G. Beyerle, R. König, C. Marquardt, T. Schmidt, L. Grunwaldt, R. Galas, T.K. Meehan, W.G. Melbourne, K. Hocke: Atmosphere sounding by GPS radio occultation: First results from CHAMP, *Geophys. Res. Lett.* **28**(17), 3263–3266 (2001)
- 38.58 J. Wickert, R. Galas, G. Beyerle, R. König, C. Reigber: GPS ground station data for CHAMP radio occultation measurements, *PCE* **26**(6–8), 503–511 (2001)
- 38.59 G.A. Hajj, E.R. Kursinski, L.J. Romans, W.I. Bertiger, S.S. Leroy: A technical description of atmospheric sounding by GPS occultation, *J. Atmos. Sol.-Terr. Phys.* **64**, 451–469 (2002)
- 38.60 W.G. Melbourne, E. Davis, C. Duncan, G. A. Hajj, K. Hardy, E. Kursinski, T. Meehan, L. Young: The application of spaceborne GPS to atmospheric limb sounding and global change monitoring Publication 94-18 (Jet Propulsion Laboratory, Pasadena 1994)
- 38.61 J. Wickert, G. Beyerle, G.A. Hajj, V. Schwieger, C. Reigber: GPS radio occultation with CHAMP: Atmospheric profiling utilizing the space-based single difference technique, *Geophys. Res. Lett.* **29**(81187), 1–4 (2002)
- 38.62 G. Beyerle, T. Schmidt, G. Michalak, S. Heise, J. Wickert, C. Reigber: GPS radio occultation with GRACE: Atmospheric profiling utilizing the zero difference technique, *Geophys. Res. Lett.* **32**(L13806), 1–5 (2005)
- 38.63 M.E. Gorbunov, S.V. Sokolovskiy, L. Bengtsson: Space refractive tomography of the atmosphere: Modeling of direct and inverse problems, Report 210 (Max Planck Institute for Meteorology, Hamburg 1996)
- 38.64 V.V. Vorob'ev, T.G. Krasil'nikova: Estimation of the accuracy of the atmospheric refractive index recovery from doppler shift measurements at

- frequencies used in the NAVSTAR system, *Phys. Atmos. Ocean* **29**, 602–609 (1994)
- 38.65 E.K. Smith, S. Weintraub: The constants in the equation for atmospheric refractive index at radio frequencies, *Proc. IRE* **41**, 1035–1037 (1953)
- 38.66 K. Hocke: Inversion of GPS meteorology data, *Annales Geophysicae* **15**, 443–450 (1997)
- 38.67 M.E. Gorbunov, S.V. Sokolovskiy: Remote sensing of refractivity from space for global observations of atmospheric parameters, Report 119 (Max Planck Institute for Meteorology, Hamburg 1993)
- 38.68 C. Marquardt, K. Labitzke, Ch. Reigber, T. Schmidt, J. Wickert: An assessment of the quality of GPS/MET radio limb soundings during February 1997, *Phys. Chem. Earth* **26**, 125–130 (2001)
- 38.69 S. Heise, J. Wickert, G. Beyerle, T. Schmidt, C. Reigber: Global monitoring of tropospheric water vapor with GPS radio occultation aboard CHAMP, *Adv. Space Res.* **27**, 2222–2227 (2006)
- 38.70 S. Healy, J. Eyre: Retrieving temperature, water vapor and surface pressure information from refractive-index profiles derived by radio occultation: A simulation study, *Quart. J. Roy. Meteorol. Soc.* **126**, 1661–1683 (2000)
- 38.71 M.A. Ringer, S.B. Healy: Monitoring twenty-first century climate using GPS radio occultation bending angles, *Geophys. Res. Lett.* **35**(L05708), 1–6 (2007)
- 38.72 G. Beyerle, T. Schmidt, J. Wickert, S. Heise, M. Rothacher, G. König-Langlo, K.B. Lauritsen: Observations and simulations of receiver-induced refractivity biases in GPS radio occultation, *J. Geophys. Res.* **111**(D12101), 1–13 (2006)
- 38.73 S.V. Sokolovskiy, C. Rocken, D. Hunt, W. Schreiner, J. Johnson, D. Masters, S. Esterhuizen: GPS profiling of the lower troposphere from space: Inversion and demodulation of the open-loop radio occultation signals, *Geophys. Res. Lett.* **33**(L14816), 1–5 (2006)
- 38.74 C.O. Ao, T.K. Meehan, G.A. Hajj, A.J. Mannucci, G. Beyerle: Lower-troposphere refractivity bias in GPS occultation retrievals, *J. Geophys. Res.* **108**(D18), 4577 (2003)
- 38.75 A.S. Jensen, M. Lohmann, H.H. Benzon, A. Nielsen: Full spectrum in version of radio occultation signal, *Radio Sci.* **38**(3), 1–15 (2003)
- 38.76 M.E. Gorbunov: Canonical transform method for processing radio occultation data in the lower troposphere, *Radio Sci.* **37**(5), 1076 (2002)
- 38.77 C.O. Ao, A.J. Mannucci, E.R. Kursinski: Improving GPS radio occultation stratospheric refractivity for climate benchmarking, *Geophys. Res. Lett.* **39**(L12701), 1–6 (2012)
- 38.78 A.J. Mannucci, C.O. Ao, L.E. Young, T.K. Meehan: Studying the atmosphere using global navigation satellites, *EOS Trans.* **95**(43), 389–390 (2014)
- 38.79 J. Wickert, G. Michalak, T. Schmidt, G. Beyerle, C.Z. Cheng, S.B. Healy, S. Heise, C.Y. Huang, N. Jakowski, W. Köhler, C. Mayer, D. Offiler, E. Ozawa, A.G. Pavelyev, M. Rothacher, B. Tapley, C. Arras: GPS radio occultation: Results from CHAMP, GRACE and FORMOSAT-3/COSMIC, *Terr. Atmos. Ocean. Sci.* **1**, 35–50 (2009)
- 38.80 S.B. Healy, J. Wickert, G. Michalak, T. Schmidt, G. Beyerle: Combined forecast impact of GRACE-A and CHAMP GPS radio occultation bending angle profiles, *Atmos. Sci. Lett.* **8**, 43–50 (2007)
- 38.81 S. Healy, A. Jupp, C. Marquardt: Forecast impact experiment with GPS radio occultation measurements, *Geophys. Res. Lett.* **32**(L03804), 1–4 (2005)
- 38.82 A.K. Steiner, B.C. Lackner, F. Ladstädter, B. Scherllin-Pirscher, U. Foelsche, G. Kirchengast: GPS radio occultation for climate applications, *Radio Sci.* **46**(RS0D24), 1–17 (2011)
- 38.83 T. Schmidt, J. Wickert, A. Haser: Variability of the upper troposphere and lower stratosphere observed with GPS radio occultation temperature-sariability of the upper troposphere and lower stratosphere observed with GPS radio occultation bending angles and temperatures, *Adv. Space Res.* **46**(2), 150–161 (2010)
- 38.84 T. Schmidt, J. Wickert, G. Beyerle, S. Heise: Global tropopause height trends estimated from GPS radio occultation data, *Geophys. Res. Lett.* **35**(L11806), 1–5 (2008)
- 38.85 A.K. Steiner, D. Hunt, S.P. Ho, G. Kirchengast, A.J. Mannucci, B. Scherllin-Pirscher, H. Gleisner, A. von Engeln, T. Schmidt, C. Ao, S.S. Leroy, E.R. Kursinski, U. Foelsche, M. Gorbunov, S. Heise, Y.H. Kuo, K.B. Lauritsen, C. Marquardt, C. Rocken, W. Schreiner, S. Sokolovskiy, S. Syndergaard, J. Wickert: Quantification of structural uncertainty in climate data records from GPS radio occultation, *Atmos. Chem. Phys.* **13**, 1469–1484 (2013)
- 38.86 S.P. Ho, D. Hunt, A.K. Steiner, A.J. Mannucci, G. Kirchengast, H. Gleisner, S. Heise, A. von Engeln, C. Marquardt, S. Sokolovskiy, W. Schreiner, B. Scherllin-Pirscher, C.O. Ao, J. Wickert, S. Syndergaard, K.B. Lauritsen, S. Leroy, E.R. Kursinski, Y.H. Kuo, U. Foelsche, T. Schmidt, M. Gorbunov: Reproducibility of GPS radio occultation data for climate monitoring: Profile-to-profile inter-comparison of CHAMP climate records 2002 to 2008 from six data centers, *J. Geophys. Res.* **117**, D18111 (2012)
- 38.87 S.P. Ho, G. Kirchengast, S. Leroy, J. Wickert, A. Mannucci, A. Steiner, C.O. Ao, M. Borsche, A. von Engeln, U. Foelsche, S. Heise, D. Hunt, B. Iijima, Y.H. Kuo, R. Kursinski, B. Lackner, B. Pirscher, M. Ringer, C. Rocken, T. Schmidt, W. Schreiner, S. Sokolovskiy: Estimating the uncertainty of using GPS radio occultation data for climate monitoring: Intercomparison of CHAMP refractivity climate records from 2002 to 2006 from different data centers, *J. Geophys. Res.* **114**, D23107 (2009)
- 38.88 A. von Engeln, S. Healy, C. Marquardt, Y. Andres, F. Sancho: Validation of operational GRAS radio occultation data, *Geophys. Res. Lett.* **36**(L1780), 1–4 (2009)
- 38.89 F. Zus, G. Beyerle, L. Grunwaldt, S. Heise, G. Michalak, T. Schmidt, J. Wickert: Atmosphere sounding by GPS radio occultation: First results from TanDEM-X and comparison with TerraSAR-X,

- Adv. Space Res. **53**(2), 272–279 (2014)
- 38.90 Status of the Global Observing System for Radio Occultation (Update 2013), IROWG/DOC/2013/02 (IROWG, 2013). <http://www.irowg.org>
- 38.91 Y.-H. Kuo, W.S. Schreiner, J. Wang, D.L. Rossiter, Y. Zhang: Comparison of GPS radio occultation soundings with radiosondes, *Geophys. Res. Lett.* **32**, 1–4 (2005)
- 38.92 W. Schreiner, C. Rocken, S. Sokolovskiy, S. Syndergaard, D. Hunt: Estimates of the precision of GPS radio occultations from the COSMIC/Formosat-3 mission, *Geophys. Res. Lett.* **34**(L04808), 1–5 (2007)
- 38.93 G.F. Lindal, G.E. Wood, H.B. Hotz, D.N. Sweetnam, V.R. Eshleman, G.L. Tyler: The atmosphere of Titan: An analysis of the Voyager 1 radio occultation measurements, *Icarus* **53**, 348–363 (1983)
- 38.94 Y.-H. Sokolovskiy, S.V. Kuo, C. Rocken, W.S. Schreiner, D. Hunt, R.A. Anthes: Monitoring the atmospheric boundary layer by GPS radio occultation signals recorded in the open-loop mode, *Geophys. Res. Lett.* **33**, L12813 (2006)
- 38.95 S.P. Ho, M. Goldberg, Y.-H. Kuo, Z.Z. Zou, W. Schreiner: Calibration of temperature in the lower stratosphere from microwave measurements using COSMIC radio occultation data: Preliminary results, *Terr. Atmos. Ocean. Sci.* **20**(1), 87–100 (2009)
- 38.96 J. Wickert, C. Arras, G. Beyerle, M. Ge, F. Flechtner, S.B. Healy, S. Heise, C.Y. Huang, B. Kuo, C. Marquardt, G. Michalak, N. Jakowski, T. Schmidt, M. Semmling: Radio occultation with navigation satellites: Recent results and prospects with Galileo, *Proc. 3rd Int. Coll. Sci. Aspects of Galileo*, Copenhagen (ESA, Noordwijk 2011)
- 38.97 L. Cucurull: Improvement in the use of an operational constellation of GPS radio occultation receivers in weather forecasting, *Wea. Forecast.* **25**, 749–767 (2010)
- 38.98 J. Aparicio, G. Deblonde: Impact of the assimilation of CHAMP refractivity profiles in Environment Canada global forecasts, *Mon. Weather Rev.* **136**, 257–275 (2008)
- 38.99 P. Poli, S. Healy, F. Rabier, J. Pailleux: Preliminary assessment of the scalability of GPS radio occultations impact in numerical weather prediction, *Geophys. Res. Lett.* **35**(L23811), 1–5 (2008)
- 38.100 F. Harnisch, S.B. Healy, P. Bauer, J. English: Scaling of GNSS radio occultation impact with observation number using an ensemble of data assimilations, *Mon. Weather Rev.* **149**, 4395–4413 (2013)
- 38.101 B. Scherllin-Pirscher, C. Deser, S.P. Ho, C. Chou, W. Randel, Y.-H. Kuo: The vertical and spatial structure of ENSO in the upper troposphere and lower stratosphere from GPS radio occultation measurements, *Geophys. Res. Lett.* **39**(L20801), 1–6 (2012)
- 38.102 A.J. Mannucci, C.O. Ao, T.P. Yunck, L.E. Young, G.A. Hajj, B.A. Iijima, D. Kuang, T.K. Meehan, S.S. Leroy: Generating climate benchmark atmospheric soundings using GPS occultation data. In: *Atmospheric and Environmental Remote Sensing Data Processing and Utilization II: Perspective on Calibration/Validation Initiatives and Strategies*, ed. by H.L. Huang, H.J. Bloom (International Society for Optical Engineering, Bellingham, WA, 630108 2006) p. 630108
- 38.103 B.C. Lackner, A.K. Steiner, G.C. Hegerl, G. Kirchengast: Atmospheric climate change detection by radio occultation data using a fingerprinting method, *J. Climate* **24**, 5275–5291 (2011)
- 38.104 B.C. Santer, T.M.L. Wigley, A.J. Simmons, P.W. Kållberg, G.A. Kelly, S.M. Uppala, C. Ammann, J.S. Boyle, W. Brüggemann, C. Doutriaux, M. Fiorino, C. Mears, G.A. Meehl, R. Sausen, K.E. Taylor, W.M. Washington, M.F. Wehner, F.J. Wentz: Identification of anthropogenic climate change using a second-generation reanalysis, *J. Geophys. Res.* **109**, D21104 (2004)
- 38.105 R. Sausen, B.D. Santer: Use of changes in tropopause height to detect human influences on climate, *Meteorologische Zeitschrift* **12**, 131–136 (2003)
- 38.106 T. Schmidt, G. Beyerle, S. Heise, J. Wickert, M. Rothacher: A climatology of multiple tropopauses derived from GPS radio occultations with CHAMP and SAC-C, *Geophys. Res. Lett.* **33**(L04808), 1–4 (2006)
- 38.107 W.J. Randel, F. Wu, W.R. Rios: Thermal variability of the tropical tropopause region derived from GPS/MET observations, *J. Geophys. Res.* **108**(D1, 4024), 1–12 (2003)
- 38.108 M. Nishida, A. Shimizu, T. Tsuda, C. Rocken, R.H. Ware: Seasonal and longitudinal variations in the tropical tropopause observed with the GPS occultation technique (GPS/MET), *J. Meteorol. Soc. Jpn.* **78**, 691–700 (2000)
- 38.109 T. Schmidt, J. Wickert, G. Beyerle, S. Heise: Global tropopause height trends estimated from GPS radio occultation data, *Geophys. Res. Lett.* **35**, L11806 (2008)
- 38.110 T. Tsuda, M. Nishida, C. Rocken, R.H. Ware: A global morphology of gravity wave activity in the stratosphere revealed by the GPS occultation data (GPS/MET), *J. Geophys. Res.* **105**, 7257–7273 (2000)
- 38.111 A. Faber, P. Llamedo, T. Schmidt, A. de la Torre, J. Wickert: On the determination of gravity wave momentum flux from GPS radio occultation data, *Atmos. Meas. Tech.* **6**, 3169–3180 (2013)
- 38.112 C. Arras, J. Wickert, C. Jacobi, G. Beyerle, S. Heise, T. Schmidt: Global sporadic E characteristics obtained from GPS radio occultation measurements. In: *Climate And Weather of the Sun-Earth System (CAWSES): Highlights from a priority program*, ed. by F.-J. Luebken (Springer, Berlin 2013) pp. 207–221
- 38.113 C.O. Ao, D.E. Waliser, S.K. Chan, J.L. Li, B. Tian, F. Xie, A.J. Mannucci: Planetary boundary layer heights from GPS radio occultation refractivity and humidity profiles, *J. Geophys. Res.* **117**(D16117), 1–18 (2012)
- 38.114 A. von Engel, J. Teixeira, J. Wickert, S.A. Buehler: Using CHAMP radio occultation data to determine the top altitude of the Planetary Boundary Layer, *Geophys. Res. Lett.* **32**(L06815), 1–4 (2005)

- 38.115 J. Wickert, E. Cardellach, M. Martín-Neira, J. Bandeiras, L. Bertino, O.B. Andersen, A. Camps, N. Catarino, B. Chapron, F. Fabra, N. Floury, G. Foti, C. Gommenginger, J. Hatton, P. Høeg, A. Jäggi, M. Kern, T. Lee, Z. Li, H. Park, N. Pierdicca, G. Ressler, A. Rius, J. Rosello, J. Saynisch, N. Soulat, C.K. Shum, M. Semmling, A. Sousa, J. Xie, C. Zuffada: GEROS-ISS: GNSS Reflectometry, Radio Occultation, and Scatterometry Onboard the International Space Station, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **9**(10), 1–30 (2016), doi:[10.1109/JSTARS.2016.2614428](https://doi.org/10.1109/JSTARS.2016.2614428)
- 38.116 E. Cardellach, F. Fabra, O. Nogués-Correig, S. Oliveras, S. Ribó, A. Rius: GNSS-R ground-based and airborne campaigns for ocean, land, ice and snow techniques: Application to the GOLD-RTR datasets, *Radio Sci.* **46**(RS0C04), 1–16 (2011)
- 38.117 C. Ruf, S. Gleason, Z. Jelenak, S. Katzberg, A. Ridley, R. Rose, J. Scherrer, V. Zavorotny: The CYGNSS nanosatellite constellation hurricane mission, *Proc. 2012 Int. Geosci. Remote Sens. Symp.*, Munich (IEEE, 2012) pp. 214–216 doi:[10.1109/IGARSS.2012.6351600](https://doi.org/10.1109/IGARSS.2012.6351600)
- 38.118 E. Cardellach, C.O. Ao, M. de la Torre Juarez, G.A. Hajj: Carrier phase delay altimetry with GPS – Reflection/occultation interferometry from low Earth orbiters, *Geophys. Res. Lett.* **31**(L10402), 1–4 (2004)
- 38.119 G. Beyerle, K. Hocke, J. Wickert, T. Schmidt, C. Marquardt, C. Reigber: GPS radio occultations with CHAMP: A radio holographic analysis of GPS signal propagation in the troposphere and surface reflections, *J. Geophys. Res.* **107**(D24, 4802), 1–14 (2002)
- 38.120 A.G. Pavelyev, A.V. Volkov, A.I. Zakharov, S.A. Kru-tikh, A.I. Kucherjavenkov: Bistatic radar as a tool for Earth investigation using small satellites, *Acta Astronaut.* **39**(9–12), 721–730 (1996)
- 38.121 C.D. Chadwell, Y. Bock: Direct estimation of absolute precipitable water in oceanic regions by GPS tracking of a coastal buoy, *Geophys. Res. Lett.* **28**(19), 3701–3704 (2001)
- 38.122 C. Rocken, J. Johnson, T.V. Hove, T. Iwabuchi: Atmospheric water vapor and geoid measurements in the open ocean with GPS, *Geophys. Res. Lett.* **32**, L12813 (2005)

Ionosphere M

39. Ionosphere Monitoring

Norbert Jakowski

Part G | 39

Global navigation satellite system (GNSS)-based monitoring of the ionosphere is important in a twofold manner. Firstly, GNSS measurements provide valuable ionospheric information for correcting and mitigating ionospheric range errors or to warn users in particular in precise and safety of life (SoL) applications. Secondly, spatial and temporal resolution of ground- and space-based measurements is high enough to explore the dynamics of ionospheric processes such as the origin and propagation of ionospheric storms.

It is discussed how ground- and space-based GNSS measurements are used to create global maps of total electron content (TEC) and to reconstruct the highly variable three-dimensional (3-D) electron density distribution on global scale under perturbed conditions. Thus, the monitoring results can be used for correcting ionospheric errors in single-frequency applications as well as for studying the driving forces of space weather-induced perturbation features at a broad range of temporal and spatial scales. Whereas large- and medium-scale perturbations affect accuracy and reliability of GNSS measurements, small-scale plasma irregularities and plasma bubbles have a direct impact on the continuity of GNSS availability by causing strong and rapid fluctuations of the signal strength, known as radio scintillations.

It is discussed how better understanding of space weather-related phenomena may help to model and forecast ionospheric behavior even under perturbed conditions. Hence, ionospheric monitoring contributes to the successful mitigation of range errors or performance degradation associated with the ionospheric impact on a broad spectrum of GNSS applications.

39.1	Ground-Based GNSS Monitoring	1140
39.1.1	Calibration of TEC Measurements	1140
39.1.2	Global Ionosphere Maps	1141
39.2	Space-Based GNSS Monitoring	1144
39.2.1	GNSS Radio Occultation	1144
39.2.2	Ionosphere/Plasmasphere Reconstruction	1145
39.3	GNSS-Based 3-D-Tomography	1147
39.3.1	Reconstruction Techniques	1147
39.3.2	Near-Real-Time Reconstruction	1148
39.4	Scintillation Monitoring	1148
39.4.1	Climatology of Radio Scintillations Deduced from GNSS Observations	1148
39.4.2	Scintillation Measurement Networks ...	1151
39.5	Space Weather	1152
39.5.1	Direct Impact of Solar Radiation and Energetic Particles	1152
39.5.2	Ionospheric Perturbations and Associated Effects	1153
39.5.3	Prediction of Space Weather Phenomena	1155
39.6	Coupling with Lower Geospheres	1156
39.6.1	Atmospheric Signatures	1156
39.6.2	Earthquake Signatures	1158
39.7	Information and Data Services	1159
	References	1159

The ionosphere is a highly variable propagation medium, mainly formed by the highly energetic part of the electromagnetic and corpuscular radiation of the Sun and its changes. Closely connected with the highly variable Sun and other geospheres, the ionosphere itself is an integral part of space weather. Complex coupling processes are not yet well understood and therefore

need further exploration. Additionally, the ionospheric impact on modern technical infrastructures relying on communication, navigation and remote sensing technologies requires permanent and reliable ionosphere monitoring and forecasting to inform users on ionospheric threats in the course of ionospheric storms according to their needs.

39.1 Ground-Based GNSS Monitoring

As pointed out in Sect. 6.3.5 of this Handbook, dual-frequency GNSS measurements can, in principle, be utilized to derive the total electron content (TEC) along the measured ray path or link. Considering their accuracy and near-real-time availability, which is permanently growing, ground-based GNSS measurements have been well established as a powerful tool for monitoring the ionosphere and related space weather effects for about two decades.

To derive TEC from ground-based Global Positioning System (GPS) measurements, various techniques have been developed [39.1–6].

Generally speaking, the procedure to derive TEC from GNSS measurements contains several steps.

In the first-order approximation, the differential phase is proportional to the integral of the electron density along the slant ray path (STEC). The differential code p and carrier phases φ can thus be written as

$$\begin{aligned}\Delta p &= p_2 - p_1 \\ &= \frac{K(f_1^2 - f_2^2)}{f_1^2 f_2^2} \text{STEC} + \Delta b_C + \Delta \varepsilon_C\end{aligned}\quad (39.1)$$

and

$$\begin{aligned}\Delta \varphi &= \varphi_1 - \varphi_2 \\ &= \frac{K(f_1^2 - f_2^2)}{f_1^2 f_2^2} \text{STEC} + \Delta b_L + \Delta \varepsilon_L.\end{aligned}\quad (39.2)$$

Here Δb_C and Δb_L designate the remaining instrumental delays including ambiguities in case of carrier phase measurements, whereas $\Delta \varepsilon_C$ and $\Delta \varepsilon_L$ designate the residual noise terms and

$$K = \frac{e^2}{8\pi^2 \varepsilon_0 m_e} \approx 40.309 \text{ m}^3 \text{ s}^{-2} \quad (39.3)$$

(Sect. 6.3.2).

Since the absolute travel time (pseudorange) measurements at coded signals are noisy due to strong multipath effects, in particular at low elevation angles, it is common practice to level the much less noisier carrier-phase measurements into the code observations (Fig. 39.1). Thereafter the leveled differential carrier-phases must be calibrated because satellite and receiver devices cause also a delay, which is commonly unknown. Instrumental biases can smoothly change, for example due to oscillator drift in the electronic circuits as a function of permanently changing conditions at satellites and receivers environments, such as the temperature. Principally there is one bias for each satellite

and for each GNSS station, which is commonly assumed to be practically constant over a period of several days. Since this calibration procedure requires some assumptions about the ionosphere, different research groups have developed different approaches.

39.1.1 Calibration of TEC Measurements

As mentioned above, the calibration requires an ionospheric model, for example simplified by a polynomial approach. Naturally, the chosen approaches are different for different research groups. Here we focus on a method that uses a well-qualified TEC model for calibration that can adequately adapt to the real ionospheric behavior [39.5–7].

The calibration approach uses the global Neustrelitz TeC model (NTCM) (Sect. 6.3.4) for TEC at epoch i according to

$$\text{STEC}_{\text{Meas}}^i = \text{STEC}_{\text{NTCM}}^i + b_{\text{RX}} - b^{\text{SAT}} + \varepsilon, \quad (39.4)$$

where $\text{STEC}_{\text{Meas}}^i$ is taken from the leveled carrier-phase measurement in (39.2), b_{RX} and b^{SAT} represent the interfrequency or differential code biases of GNSS receiver and corresponding GNSS satellite respectively. The term ε explicitly represents the residual errors. The term $\text{STEC}_{\text{NTCM}}$ stands for slant TEC derived from the used NTCM model that depends on typical geophysical parameters and current solar activity level approximated by the solar radio flux index F10.7 [39.7]. Since the TEC model provides geometry-free vertical TEC, the required $\text{STEC}_{\text{NTCM}}$ must be converted from ver-

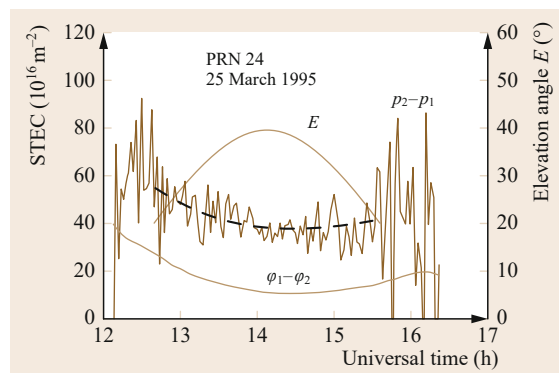


Fig. 39.1 Differential code ($p_2 - p_1$) and carrier ($\varphi_1 - \varphi_2$) phases of GPS signals (PRN24) measured in Neustrelitz on 25 March 1995. The *dashed line* indicates the least squares leveled differential carrier-phases used for further computation (after [39.5])

tical to slant TEC using an elevation (E)-dependent obliquity factor or mapping function $M(E)$ as considered in Sect. 6.3.4.

Now (39.4) can be rewritten as

$$\text{STEC}_{\text{Meas}}^i = M(E) \text{VTEC}_{\text{NTCM}}^i + b_{\text{RX}} - b^{\text{SAT}} + \varepsilon \quad (39.5)$$

in which the local TEC model values at ionospheric piercing points are directly used for the subsequent calibration procedure. The model approach allows for the choosing of an autonomous model like NTCM with fixed coefficients or an operational model NTCM_{op} with reduced dependencies and variable coefficients, which better fits the actual ionospheric conditions. If data coverage is sufficient, the latter method is recommended, whereas in the case of poor data coverage the autonomous model should yield best results.

The model coefficients and interfrequency biases (IFB) b_{RX} and b^{SAT} are then obtained by weighted least squares fit of (39.5) to the observation data at all measured radio links for a previous 24 h period [39.5].

Since the instrumental biases in (39.5) cannot be separated, and to find a comparable measure of satellite biases, the common practice is to introduce an additional equation that defines

$$\sum_i b^{\text{SAT}_i} = 0 \quad (39.6)$$

for all satellites and measurements.

Thanks to the model-assisted calibration technique, the calibration of IFBs and the subsequent generation of TEC maps can be carried out in near-real time. A satellite bias sample of GPS satellite G07 monitored over some days in October 2010 is shown in Fig. 39.2.

39.1.2 Global Ionosphere Maps

The ionosphere may cause signal delays in GNSS observations that correspond to link-related range errors of up to 100 m. Whereas this error can mostly be corrected in dual-frequency measurements by a linear combination of these two measurements, single-frequency measurements need additional information to mitigate the ionospheric error. Knowing that the first-order range error is proportional to TEC, it becomes evident that ionospheric corrections can be provided by a TEC model or by TEC maps, for example deduced from corresponding GNSS measurements. Whereas current TEC models provide only climatological data, actual TEC maps can provide more realistic correction data. The latter method is typically used in satellite-based

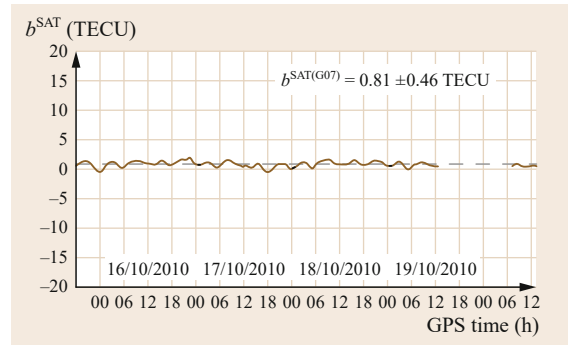


Fig. 39.2 Sample of bias computation results on a few selected days in October 2010 providing a satellite bias of $b^{\text{Sat(G07)}} = 0.81$ TECU at a root mean squared (RMS) uncertainty of 0.46 TECU (1 TECU = 10^{16} electrons/m²)

augmentation systems (SBAS) such as **WAAS**, **EGNOS**, **GAGAN** or **MSAS**. Here a number of reference stations provide the basic information for computing regional TEC maps. Such a map provides TEC values at a regular grid with spacings of $5^\circ \times 5^\circ$ in latitude and longitude respectively. It is obvious that the accuracy and spatial resolution of generated TEC maps depends on the availability of ground-based GNSS measurements over the area in view.

As noted earlier, different approaches have been developed by different research groups. *Jakowski et al.* [39.5, 6], for instance, use a model-assisted technique for calibrating and mapping the measured TEC over Europe since 1995, over the polar areas since 2001 and globally since 2010.

TEC maps are generated by assimilating available observations into a specific background model [39.7] as considered in Sect. 6.3.4. This procedure has the advantage that even in the case of poor or uneven data coverage, for example over the oceans, climatological TEC estimations are available for range error corrections by customers.

The TEC data assimilation procedure starts with a least squares adjustment of the background model with respect to all observations. Assuming that the adjustment value at measurement epoch i is $\text{VTEC}_{\text{adj}}^i$, the deviations at all N ionospheric piercing points j are computed according to

$$\Delta \text{VTEC}_j^i = \text{VTEC}_j^i - \left(\text{VTEC}_{\text{NTCM}_j}^i + \text{VTEC}_{\text{adj}}^i \right). \quad (39.7)$$

The model adjustment takes care that the sum of positive and negative deviations of all N observations j from

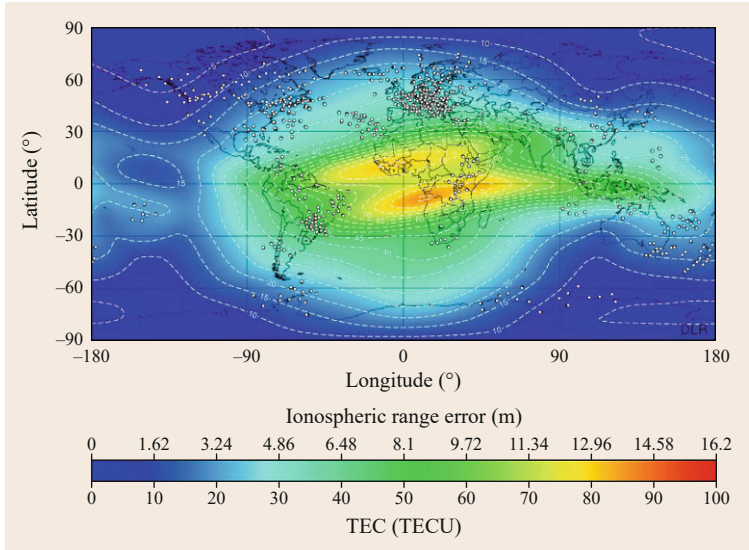


Fig. 39.3 Sample VTEC map on 17 October 2011 at 14:00 UT. Ionospheric piercing points of GPS ground stations are marked by dots

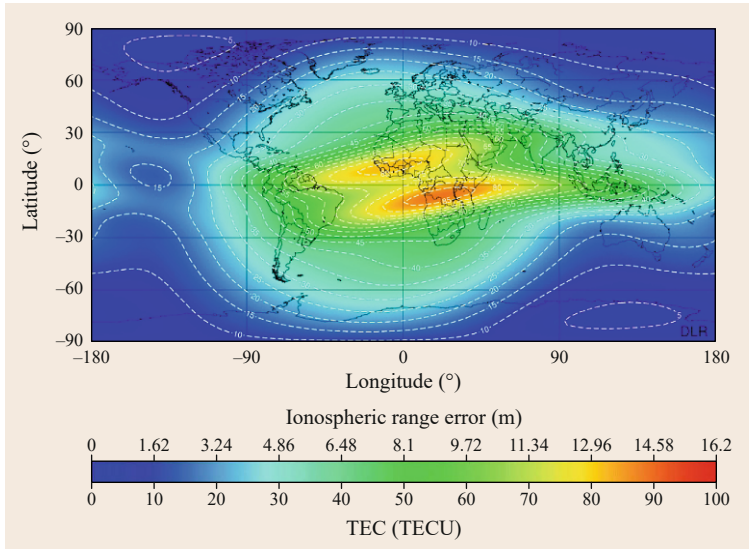


Fig. 39.4 Map of the VTEC background model corresponding to the VTEC map shown in Fig. 39.3

the model surface is equal to zero

$$\sum_{j=1}^N \Delta \text{VTEC}_j^i = 0. \quad (39.8)$$

ΔVTEC values at grid points (GP) (k, l) are simply computed by summing up weighted deviations of surrounding piercing points. The weight decreases with the distance using a Gaussian-type weighting function. The width of the weighting function is a tuning parameter that defines the influence of individual piercing-point values in the sense of a correlation length. So for instance the parameter can be set small at high data density and big at poor data coverage. To reduce mapping

errors at low elevation angles, an elevation-dependent weighting function can be added. The **VTEC** (vertical total electron content) values at grid points $\text{GP}(k, l)$ are finally computed as

$$\begin{aligned} \text{VTEC}^i(k, l) = & \Delta \text{VTEC}^i(k, l) + \text{VTEC}_{\text{NTCM}}^i(k, l) \\ & + \text{VTEC}_{\text{adj}}^i. \end{aligned} \quad (39.9)$$

The final results represent measured TEC values in the vicinity of the piercing points, whereas at greater distances from measurements somewhat modified model values are derived. Following this procedure, VTEC

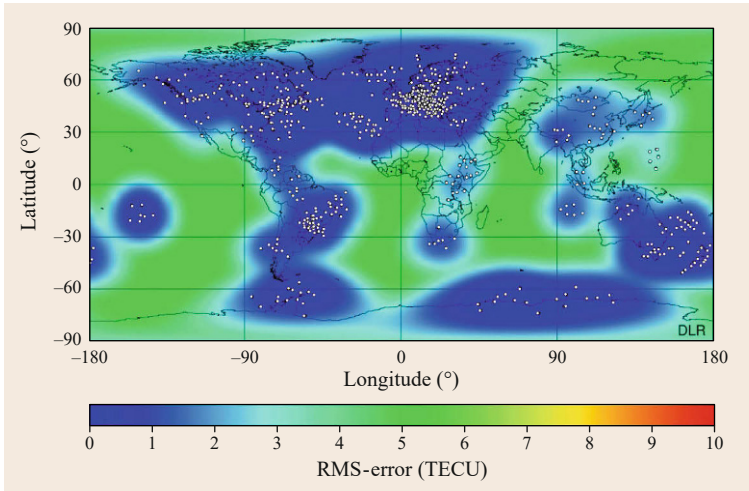


Fig. 39.5 Error map of VTEC that corresponds to the generation of the sample VTEC map shown in Fig. 39.3

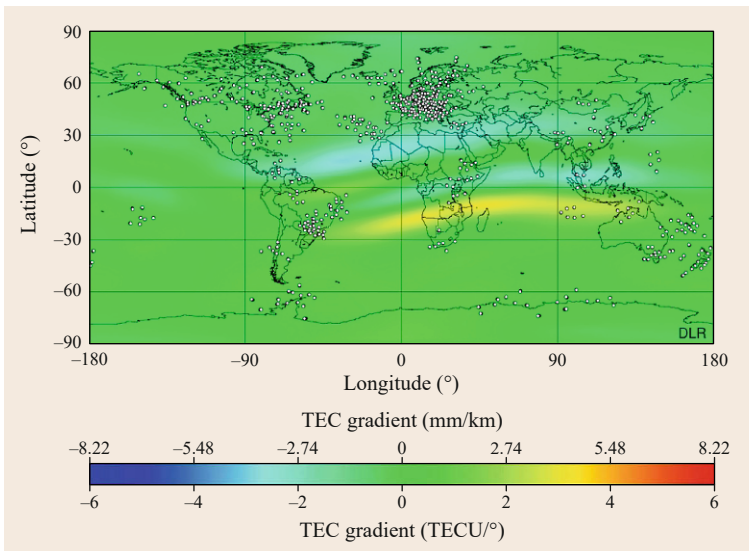


Fig. 39.6 Map of latitudinal gradients of VTEC corresponding to the VTEC map shown in Fig. 39.3

maps and related products have generated in the Space Weather Application Center (SWACI) at DLR Neustrelitz for many years [39.8].

A global sample map of vertical TEC is shown in Fig. 39.3 for illustration. Most data originate from the geodetic network of the international GNSS service (IGS) network [39.9]. Related piercing points are indicated in the map by small circles. Grid values are spaced by $2.5^\circ \times 5^\circ$ in latitude–longitude. The corresponding map of the operational background model NTCM_{op} is shown in Fig. 39.4.

Since the difference between the model and the derived TEC map is obviously not very big, the model approach fits quite well to the ionosphere in this geomagnetically quiet day. This can be checked when inspecting the corresponding error map as shown in

Fig. 39.5. The residual range error is in the order of less than 1 TECU (16.2 cm at L1 GPS frequency) in those regions where good data coverage exists. Over regions not covered by data, for example over the oceans, TEC remains uncertain to a higher degree related to the applied model.

Once TEC maps have been derived from observational data, a multitude of secondary information can be derived such as TEC gradient or rate-of-TEC (RoT) maps. Taking into account also the high temporal resolution of 1 s used in operational processing, perturbation processes can be monitored in detail as considered in Sect. 39.5. Since TEC gradient information is important for precise positioning and navigation and safety of life applications [39.10, 11], TEC gradients as shown in Fig. 39.6 can be used for a first-order estimate of po-

tential problems that might be caused by the handling of ionospheric gradients in an operational system.

The derived TEC gradients are rather small compared with those in the order of 100 mm/km and more, which might cause problems in SoL applications. TEC

map-derived gradients are smooth due to the fact that only large-scale phenomena are imaged at grid spacings of more than 100 km. Thus, for deriving TEC gradient-induced ionospheric threats, original slant TEC measurements must be analyzed [39.11].

39.2 Space-Based GNSS Monitoring

Space-based dual-frequency GNSS measurements use the same observation equations (39.1) and (39.2) for deriving ionospheric information as used in ground-based techniques.

GNSS measurements on board low Earth orbiting (LEO) satellites may essentially contribute to monitoring of the geoplasma. Thus, GNSS radio occultation measurements are capable of monitoring the vertical ionization of the ionosphere on global scale [39.12–14]. Additionally, regular navigation data used for satellite positioning can effectively be utilized to monitor the three-dimensional electron density distribution of the topside ionosphere/plasmasphere near the orbit plane [39.15]. The effectiveness of radio occultation measurements has been demonstrated by several satellite missions such as Microlab-1 with the GPS/MET experiment [39.12], CHAMP [39.14] and in particular by the COSMIC/Formosat-3 mission in recent years [39.16].

39.2.1 GNSS Radio Occultation

GNSS radio occultation is a limb-sounding technique that enables the retrieval of the vertical refractivity profile of planetary atmospheres. What is measured is the ray path bending and/or the phase of the radio wave while approaching the planetary surface in the limb-sounding geometry.

Radio occultation techniques were successfully applied for exploring planetary atmospheres from Mars and Venus using Mariner IV and Venera 4 signals respectively. In the late 1980s Yunck et al. [39.17] proposed applying the radio occultation techniques also to the Earth's atmosphere sounding using the L-band signals of GPS, which has just been established. The GPS/MET experiment on board the Microlab 1 satellite mission that was launched in April 1994 has convincingly demonstrated that the GPS radio occultation technique is a powerful tool for remote sensing of the Earth's neutral atmosphere and ionosphere [39.12]. A few years later other satellite missions such as SAC-C, Oerstedt, CHAMP and GRACE followed to exploit the new technique for exploring the lower and middle atmosphere and the ionosphere. A breakthrough was

the launch of six LEO satellites (COSMIC/Formosat-3 mission) in 2006 [39.16]. Whereas a single-satellite mission like CHAMP can provide about 200–400 measurements, the multisatellite COSMIC/Formosat-3 mission can provide up to ≈ 2500 occultations per day.

The radio occultation geometry is schematically shown in Fig. 39.7. The refraction angle α between ray path asymptotes of LEO and GNSS satellites has to be derived from GNSS carrier-phase measurements on board the LEO satellite with high accuracy. Since the bending angle is principally less than one degree, the orbit data must be measured precisely (centimeter range) and clock drifts have to be corrected. The refraction angle α is related to the refraction index n via the integral equation

$$\alpha(a) = -2a \int_{r_0}^{\infty} \frac{1}{\sqrt{r^2 n^2 - a^2}} \frac{d \ln(n)}{dr} dr. \quad (39.10)$$

Here the impact or approaching parameter a describes the refractive distance of the asymptotic ray paths from the center of the Earth. The vertical refractive index profile is then retrieved as a function of α and a by applying the Abel integral transform [39.18] to integrate (39.10).

Considering only the upper part in ionospheric heights and the ionospheric range error formula (6.85) in Chap. 6, it becomes clear that the Abel inversion

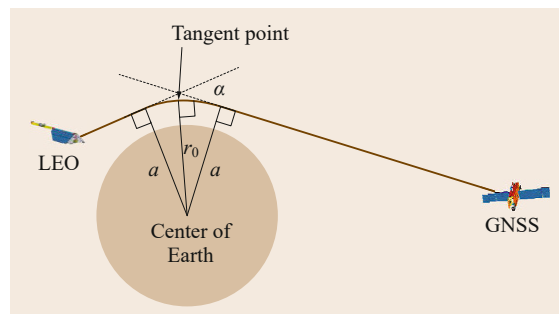


Fig. 39.7 GNSS radio occultation geometry for retrieving vertical refractivity profiles in the troposphere and ionosphere

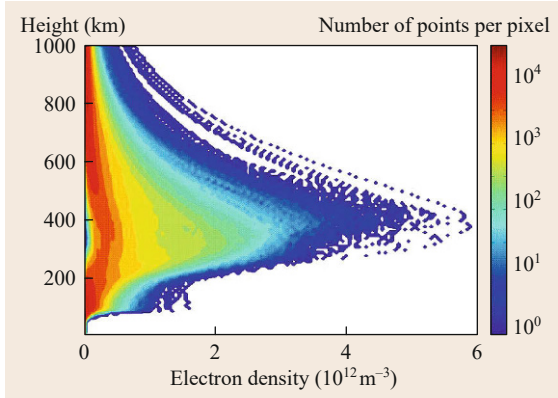


Fig. 39.8 Superposed plots of 30 000 vertical electron density profiles obtained from IRO measurements on board CHAMP in 2002, a year of high solar activity (after [39.14])

of (39.10) provides the vertical electron density profile from the satellite orbit height down to the bottom side ionosphere. In the lower atmosphere the vertical refractive index profile reveals the neutral gas temperature in conjunction with the water vapor profile [39.19]. Due to the dispersive nature of the ionosphere, *ionospheric radio occultation* (IRO) measurements can directly use differential GNSS carrier-phases from dual-frequency GNSS measurements on board the LEO satellite for retrieving vertical electron density profiles.

Assuming a commonly used sampling rate of 1 Hz, the height resolution is in the order of a few kilometers. Accuracy can be improved by taking into account ray path bending effects [39.20, 21]. Continuous IRO measurements on satellites such as CHAMP, COSMIC or GRACE allow for the collection of massive datasets well suitable for ionospheric studies (Fig. 39.8) and for developing and/or improving models of ionospheric key parameters such as the peak density N_mF2 [39.22] or the peak height h_mF2 of the F2 layer [39.23].

Averaged IRO measurements show typical ionospheric features such as the equatorial anomaly and the mid-latitude or main trough at nighttime as illustrated in Fig. 39.9.

The worldwide distribution of the electron density indicates a striking geomagnetic control of the F2 layer [39.24]. The most distinctive features are the so-called crest regions of maximum electron density appearing at both sides of the geomagnetic equator at a distance of about 15° , known as the equatorial anomaly. The equatorial anomaly is usually more pronounced around noon than around midnight. The ionospheric mid-latitude or main trough [39.25] is a zone of low electron density located between about 50° and

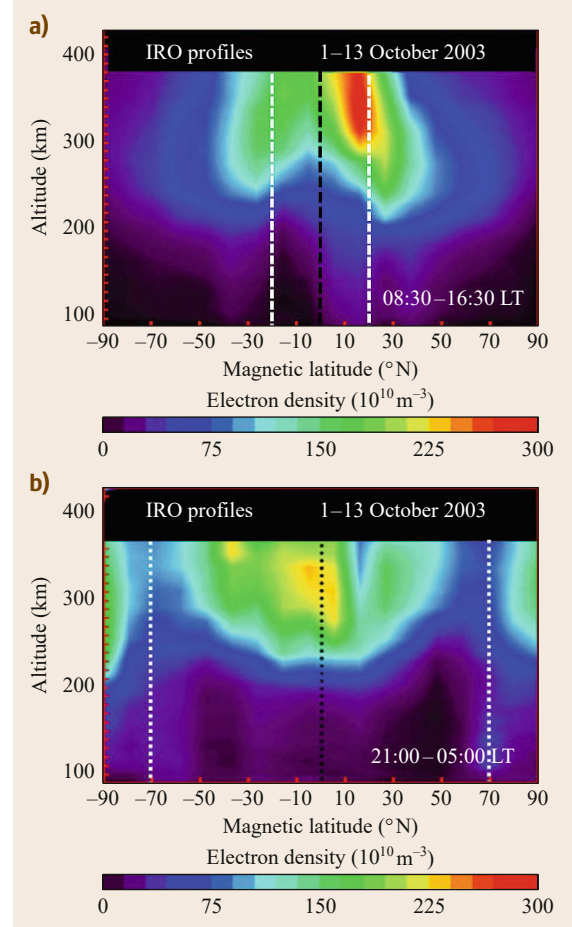


Fig. 39.9a,b Averaged IRO profiles retrieved from 1–13 October 2003 at daytime (a) and at nighttime (b). The geomagnetic equator is marked by a black dashed line, whereas crest and trough latitudes are marked by white dashed lines

70° geomagnetic latitude. The main trough lasts from about 18:00 to about 06:00 magnetic time and its width is about 500–1000 km [39.26]. The electron density inside the trough is drastically reduced by as much as a factor of two at 1000 km height and as much as an order of magnitude at the F2 layer peak height.

39.2.2 Ionosphere/Plasmasphere Reconstruction

Dual-frequency GNSS data measured on board LEO satellites for positioning can effectively be utilized for ionospheric monitoring in the same way as ground-based measurements. The moving receiver on board the LEO satellite enables the collection of numerous measurements during one satellite revolution. Usually

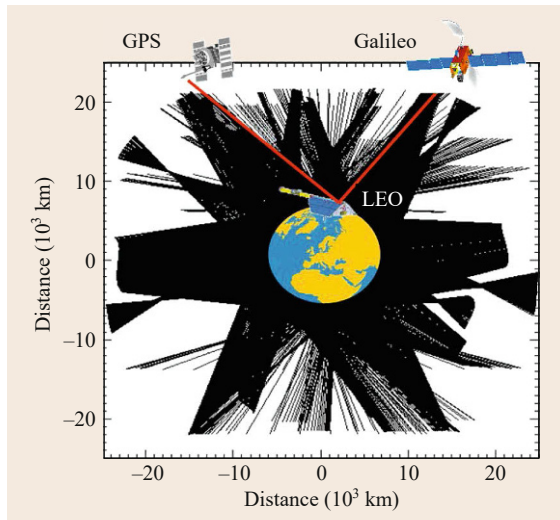


Fig. 39.10 Illustration of the distribution of radio links between GNSS and a LEO satellite in the LEO orbit plane during one satellite revolution

the data are not homogeneously distributed as demonstrated in Fig. 39.10.

One satellite revolution enables up to about 4000 GPS measurements, a number that would increase if signals from other GNSSs could be used in addition. In a first-order approach the ionosphere is assumed to be static during one revolution (93 min). In agreement with the available data coverage this assumption is justified for monitoring large-scale phenomena in the electron density distribution near the satellite orbit plane. To overcome problems related to inhomogeneous data distribution and data gaps as seen in Fig. 39.10, the reconstruction of the electron distribution can be realized in a reasonable way by assimilating link-related calibrated TEC data into a reliable background model. Following [39.15] the spatial electron density distribution is numerically defined in a specific voxel structure, which is initialized by a model, here by the parameterized ionospheric model (PIM) [39.27]. To meet all link-related TEC measurements, an iterative process is carried out that modifies the electron density inside the voxels crossed by the CHAMP-GPS radio links until the residuals are less than a certain cutoff level. The iterative procedure resembles the well-known multiplicative algebraic reconstruction technique (MART).

To illustrate the achieved results, Fig. 39.11 shows a typical reconstruction of the electron density distribution from CHAMP to GPS orbit heights in the CHAMP orbit plane.

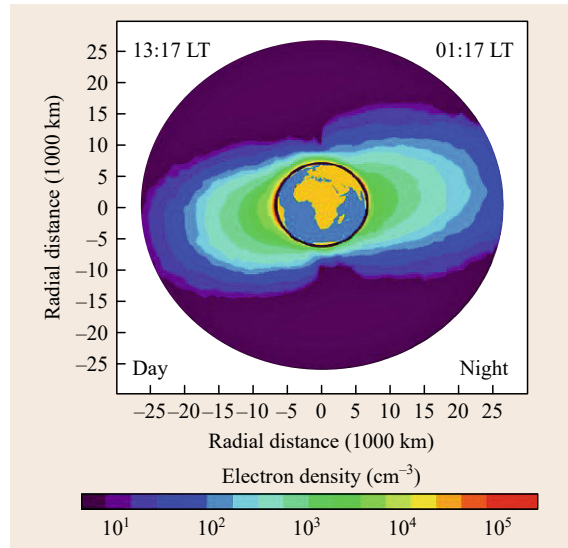


Fig. 39.11 Reconstruction of the electron density distribution of the topside ionosphere based on GPS data received on board CHAMP. The reconstruction is generated from medians at 21:00 UT over ten consecutive days in August 2005. The *right side* shows the ionosphere/plasmasphere shortly after midnight, whereas the *left side* represents the ionosphere shortly after noon

Although the reconstruction will not be an absolutely correct reproduction of the real state of the topside ionosphere and plasmasphere, the result is an improved model output. Independent of data coverage the result is always stable and physically reasonable. The assimilation process results in a three-dimensional electron density distribution near the CHAMP orbit plane. The associated distribution in the orbit plane is seen in Fig. 39.11. It is obvious that the global view on the three-dimensional structure of Earth's plasma environment enables the study of magnetospheric-ionospheric coupling processes.

In Fig. 39.11 the compression of the plasmasphere due to the solar wind is nicely seen at day-side, whereas the night-side clearly shows an enlarged extension of the plasmasphere. Since the field-aligned structure of this geoplasma is sensitive to space weather changes it becomes evident that space-based GPS measurements can essentially contribute to space weather monitoring and studying ionosphere-magnetosphere coupling processes. To enhance the spatial resolution, a multi-LEO satellite constellation capable of using signals from different GNSSs would be beneficial in the future.

39.3 GNSS-Based 3-D-Tomography

The three-dimensional reconstruction of the electron density distribution of the full global ionosphere and plasmasphere systems is a challenging task for upcoming years. Only the knowledge of the three-dimensional electron density distribution will provide sufficient insight for the understanding of physical processes behind the observed dynamics of the ionosphere. Additionally, reliable high-frequency (HF) communications and accurate knowledge of ionospheric errors in GPS-based navigation systems are much needed by customers. In GNSS practice, due to the 2-D simplification of the 3-D ionosphere, TEC information may be insufficient in particular in regions of strong horizontal gradients of ionization. The growing availability of GNSS data will improve the conditions for 3-D or 4-D reconstruction of the electron density dramatically. The number of data increases due to densification of ground-based GNSS networks, due to the availability of more and more space-based GNSS data from satellite missions and due to the availability of several GNSS systems like GPS, [GLONASS](#), Galileo or Beidou/Compass. As in case of TEC mapping the reconstruction of the 3-D electron density distribution of the ionosphere and plasmasphere systems in near-real time or postprocessing will benefit from ground- and space-based measurements considered in previous sections.

39.3.1 Reconstruction Techniques

Since the first idea thought up by *Austen* et al. in 1986 [39.28] to apply tomographic techniques to the imaging of the ionospheric electron distribution, various methods have been developed to reconstruct the electron distribution from integral TEC measurements. In the early years of ionospheric tomography, TEC from dual-frequency beacon measurements was used [39.29]. Receiving beacon signals, for example those of the forerunner of GPS, the Navy Navigation Satellite System (NNSS), along a chain of ground receivers, multiple intersecting measurements can be made for imaging the vertical electron density along the receiver chain for a selected satellite pass. When GPS signals became available, the imaging was not spatially restricted to a plane defined by the satellite orbit in conjunction with the beacon receiver chain and not restricted in time due to the permanent availability of at least a few GPS satellites needed for positioning [39.30]. A common geometrical constraint exists both for beacon as well as for ground-based GNSS measurements. This is the lack of horizontal measurements causing an uncertainty in the vertical structure of the resulting image. To overcome this ill-posed dataset in

order to get stable solutions, one has to include some constraints in the reconstruction procedure, which may differ and lead to a wide range of approaches that cannot be considered here.

The inverse problem of ionospheric tomography is the following: for a given set of path integrals of electron density (TEC) along known transionospheric paths, find the electron density distribution that satisfies all measurements in an optimal way.

The first reconstructions by *Austen* et al. [39.28] used the iterative algorithm of the algebraic reconstruction technique (ART) to image intersecting TEC measurements. Here the initial state is modified by adding corrections that are a function of the difference between the measured TEC and the initial TEC. The solution is achieved after iterating the algorithm several times until the result converges. The multiplicative algebraic reconstruction technique (MART) is similar to ART, however, in this case, the ionospheric state is modified with a factor for correction instead of adding a correction term. This approach ensures that the relative electron density along the ray path remains unchanged, i.e., the electron density cannot become negative. This type of reconstruction technique has been applied in principle for imaging the 3-D topside ionosphere/plasmasphere electron density distribution in the vicinity of the CHAMP orbit plane considered in the previous section [39.15]. Here the initial information is provided by the parameterized ionospheric model (PIM) [39.27] in which the data are assimilated.

Most of the powerful imaging approaches that have been developed in recent years are assimilative since they combine actual measurements with background information provided by an empirical or physical model or at least by empirical orthonormal functions derived from Chapman profiles [39.31–36]. Assimilation techniques commonly use Kalman filtering and 3-D variational techniques [39.37].

Some of these techniques are the ionospheric data assimilation three dimensional (IDA3-D) developed by *Bust* et al. [39.31, 32] the multi-instrument data analysis software (MIDAS) developed by *Mitchell* and *Spencer* [39.34] and physics-based techniques such as the global assimilation ionospheric model (GAIM), developed by *Schunk* et al. [39.33] at Utah State University, and *JPL-GAIM*, developed by the Jet Propulsion Laboratory (JPL) and the University of Southern California [39.36].

The electron density assimilative model (EDAM) has been developed by *Angling* and *Cannon* at QinetiQ (UK) [39.35] to assimilate ionospheric measurements, in particular ground- and space-based GNSS measure-

ments, into the IRI 2007 model [39.38]. The background model is modified by the difference between the observation vector and the observation operator, which links the geometry of observations to the background model. Before the update starts, this difference is scaled with a weight matrix, which is composed of the error covariance matrices of the background model and the observables. The assimilation is based on a form of minimum variance optimal estimation, also referred to as best linear unbiased estimation (BLUE) that provides an expression for an updated estimation of the ionospheric state. Using the background model, electron densities are established on a 3-D grid of voxels in a magnetic Sun-fixed coordinate system.

MIDAS (multi-instrument data analysis software) is a three-dimensional, time-dependent algorithm for imaging the ionosphere. MIDAS uses phase data from dual-frequency GNSS data to measure relative TEC between satellites and ground receivers. As a priori information a set of orthonormal profiles derived from Chapman functions are used to constrain the vertical profile, so only a small number of coefficients is needed to estimate electron densities above a selected geographic location. The horizontal distribution is determined by a spherical harmonic expansion. Since MIDAS uses differential carrier-phase data for estimating biased TEC, the procedure requires a time-dependent algorithm to gain ionospheric information from ionospheric changes when GNSS satellites move across the sky. Thus, MIDAS can be considered as being a 4-D imaging technique.

The first-principle ionospheric physics-based background model global assimilative ionospheric model (GAIM) of JPL is a global, fully three-dimensional, and time-dependent ionospheric model. It numerically solves for ion and electron densities through the hydrodynamic equations for ions and incorporates state-of-the-art Kalman filter and four-dimensional variational (4DVAR) approaches that enable the assimilation of various types of ionospheric measurements.

The space weather model *global assimilation of ionospheric measurements* (GAIM) developed by the Utah State University team at the Space Weather Center, is similar to the tropospheric weather models run by NOAA. This model provides real-time specifications and forecasts for global distributions of upper atmosphere/ionosphere densities, temperatures, and winds. The GAIM space weather model originally became an operational Air Force model in December 2006. The data types currently being used in GAIM include line-of-sight TEC measurements deduced from ground-based GPS receiver networks and space-borne GPS receivers, satellite ultraviolet (UV) limb scans, and ionosonde data. Intensive validation has also been conducted using various independent data sources, including vertical TEC measured using satellite ocean altimeter radars (such as those aboard TOPEX and Jason-1 missions), ionosondes, and incoherent scatter radars.

39.3.2 Near-Real-Time Reconstruction

Whereas ionospheric research doesn't require near-real-time imaging of the ionospheric electron density, precise GNSS applications benefit from accurate 3-D images to solve phase ambiguities in near-real time. The availability of 3-D images avoids mapping errors due to serious simplifications of the vertical ionospheric structure in TEC-based ionospheric delay estimates. However, the 3-D reconstruction is only beneficial when the reconstruction is sufficiently accurate. Since temporal and spatial resolution in conjunction with the accuracy of 3-D images depends to a great extent on the imaging technique and on the density of reliable data, further investigation is required to fulfill challenging customer needs. Nevertheless, several attempts are currently being made to provide three-dimensional electron density reconstructions in near-real time that still need comprehensive validation via international co-operation (Sect. 39.7).

39.4 Scintillation Monitoring

Radio scintillations as discussed in Sect. 6.3.3 impact in particular the availability of GNSS signals, for example for positioning, navigation and time transfer (PNT) services. Hence, any signal failure could cause serious problems in complex infrastructures. Severe scintillations may cause loss of lock of GNSS signals [39.39] thus reducing the availability of GNSS services. Based on a better understanding of the physics of the ionospheric plasma and its irregularities, forecast tools shall be developed to forecast the onset of scintillation activ-

ity 3–6 h in advance. The estimation of the scintillation probability up to several hours in advance is an important issue to enhance the reliability of GNSS.

39.4.1 Climatology of Radio Scintillations Deduced from GNSS Observations

The S_4 index introduced in Sect. 6.3.3 as a well-accepted measure of radio scintillation activity depends strongly on geophysical conditions, such as local time,

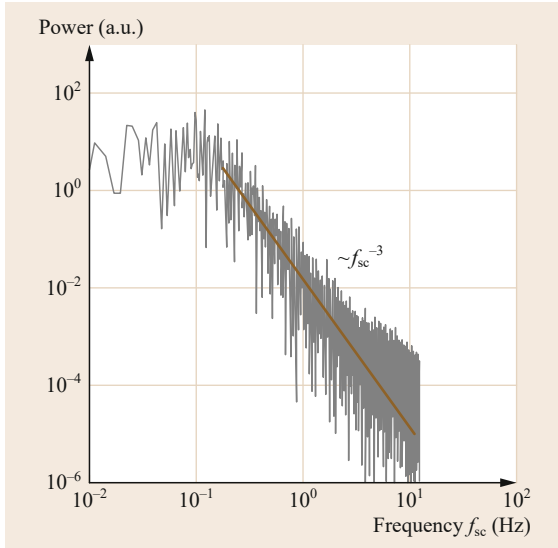


Fig. 39.12 Spectral power density of GNSS amplitude fluctuations measured on 5 April 2006 in Bandung, Indonesia

season, latitude, and solar activity level [39.40, 41]. Since the fluctuation rate of typical GNSS scintillations is up to 10 Hz and more, GNSS-based scintillation monitoring for scientific analysis requires well-qualified GNSS receivers with high sampling rate in the order of 50 Hz or more (Fig. 39.12). In principle all high-rate GNSS receivers providing access to the signal intensity can be used in the measurement praxis. It is worth noting that a broad science community uses a special scintillation receiver (Novatel GSV 4000) for scintillation monitoring.

Long-term studies on scintillations have shown a strong dependence on solar activity and geophysical conditions such as location, season and daytime. As already pointed out in Sect. 6.3.3, scintillations can primarily be observed at high and low latitudes where they are generated by different physical processes. Scintillation effects observed at a high-latitude station are shown in Fig. 39.13.

In the auroral and polar cap latitudes, any significant magnetic storm activity can produce scintillation effects. Severe amplitude and phase scintillations have been observed in coincidence with steep TEC gradients, a characteristic of the edge of polar cap patches. Such gradients may result in the production of small-scale irregularities caused by the gradient-drift instability [39.42, 43]. Due to this storm-driven behavior there is no clear dependence on local time reported. Furthermore, it is a well-known fact that high-latitude scintillations are not as severe as those measured in the near-equatorial belt. However, high-latitude scintil-

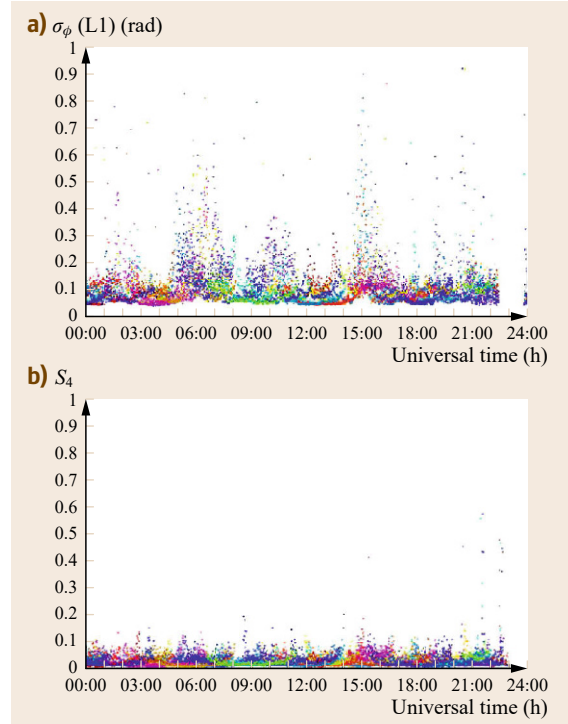


Fig. 39.13a,b S_4 and σ_ϕ indices observed at Kiruna station (67.84°N; 20.41°E) on (a) 7 and (b) 8 March 2012

lations can last for many hours, even days. The observation results shown in Fig. 39.13 confirm the statement that phase scintillations are much more pronounced than signal-strength fluctuations at high latitudes. This indicates a predominance of refractive index fluctuations Δn versus diffraction or scattering effects in this region.

The maximum fading depth observed on GPS signals from the north polar cap region was approximately 10 dB, whereas in the equatorial anomaly region the fading depth is observed as much as 25 dB [39.44]. In general the peak-to-peak fading depth can be estimated by the formula [39.45]

$$P_{\text{Fluc}} = 27.5 S_4^{1.26}. \quad (39.11)$$

At low latitudes the occurrence probability of scintillations is more regular due to solar illumination-driven thermosphere/ionosphere coupling processes. Thus, significant enhancements of the S_4 index can be observed at low latitudes regularly after sunset except during low solar activity conditions (Fig. 39.14).

GNSS measurements at the low-latitude station at the Bahir Dar University, Ethiopia, show a strong enhancement of signal strength fluctuations immediately after sunset that is surely associated with the Rayleigh-

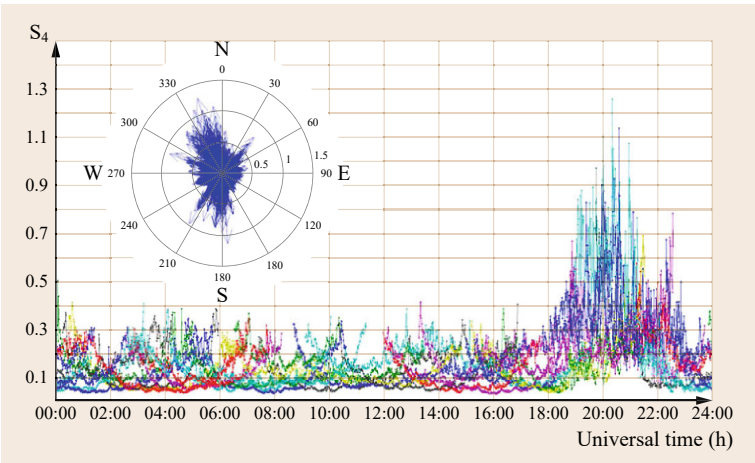


Fig. 39.14 S_4 scintillation measurements at Bahir Dar, Ethiopia (11.6° N; 37.4° E) on 14 April 2012 made on all available satellites (color marked). Onset of enhanced activity starts after sunset

Taylor plasma instability. The increased S_4 activity lasts typically a few hours until midnight (Fig. 39.14). As expected, scintillation activity is most pronounced in the north-south direction because Bahir Dar at (11.5° N; 37.4° E) is located between the northern and southern ionospheric crest over Africa.

Long-term observations of scintillation activity allow estimation of the occurrence probability of scintillations besides scintillation strength or fading depth.

Although the solar activity was very low in 2006, scintillations have been observed in Bandung at an occurrence probability of about $6.0 \cdot 10^{-3}$ for severe scintillations with $S_4 \geq 0.8$. Long-term studies shall further improve the reliability of occurrence statistics. Although the seasonal dependence usually shows maxima around the equinoxes at the Asian sector as in Bandung, this is not valid at the American low-latitude sector, where scintillation probability is higher during solstices than during equinoxes.

It is worth mentioning that signal-strength fluctuations are closely associated with rapid phase changes of TEC at low latitudes. This relationship has been tested by a direct comparison of simultaneous S_4 and rate-

of-TEC (RoT) measurements in Bandung [39.46] as Fig. 39.15 shows.

TEC rate measurements are effective in detecting plasma bubbles whose occurrence is associated with enhanced scintillation activity as described in [39.47, Sect. 6.3].

TEC depletions occur when one or more plasma bubbles drift across the line of sight between the GPS receiver and the satellite. Therefore, sudden TEC depletions in link-related dual-frequency GNSS measurements can be considered as a manifestation of equatorial plasma bubbles. Occurrence characteristics of plasma bubbles derived from global ground-based GPS receiver networks were studied by *Nishioka et al.* [39.48]. The standard deviation of RoT, which is commonly known as rate of TEC change index (ROTI), is used to identify small-scale fluctuations. ROTI is often used to investigate ionospheric fluctuations [39.49, 50]. In analogy to the low-latitude S_4 behavior the diurnal variation of ROTI shows a strong enhancement in the evening hours between sunset and midnight.

Large-scale structures associated with ionospheric storms may also cause large TEC depletions resulting

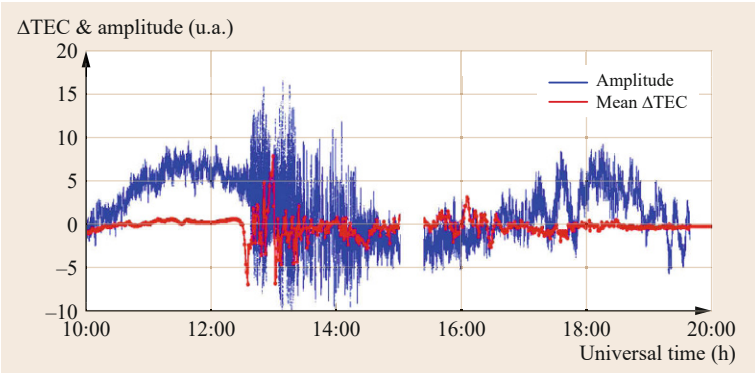


Fig. 39.15 Comparison of signal amplitude (blue) and TEC rate (red) measured simultaneously in Bandung, Indonesia on 5 April 2006 (after [39.46])

in enhanced scintillation activity as found over Africa during a moderate geomagnetic storm on 24–25 October 2011 [39.47]. High scintillation activity may lead to a loss of lock of the GNSS signals. To avoid problems in accurate and safety-critical applications, a more comprehensive understanding of physical processes behind radio scintillations is still needed for developing mitigation techniques and forecast tools. International initiatives encouraging systematic and worldwide monitoring of ionospheric irregularities are very helpful on this matter.

39.4.2 Scintillation Measurement Networks

Since space weather impact on L-band signals of GNSS cannot be ignored, tremendous international efforts are underway to measure and model ionospheric irregularities and resulting scintillation effects at radio signals. As pointed out in the previous section, GNSS measurements essentially help to collect representative datasets for comprehensive theoretical and empirical data analysis.

Whereas the US Air Force Research Laboratory (AFRL) is establishing the low latitude SCINDA (SCINtillation and Decision Aid) ground station network [39.51], the European Space Agency (ESA) is supporting the global deployment of scintillation receivers within the PRIS and MONITOR projects [39.52, 53].

In addition to SCINDA the US Air Force Research Lab has initiated the launch of the communication/navigation outage forecast system (C/NOFS) satellite in 2008 [39.54]. The main goal is developing forecast tools for scintillation occurrence probability. The estimation of the scintillation probability up to several hours in advance is an important issue to enhance the reliability of GNSS.

Besides ESA activities in supporting coordinated scintillation monitoring in regional and global networks, several European countries have been active in establishing national GNSS networks, contributing to improving our knowledge on scintillations. Among them are a high-latitude network operated by Istituto Nazionale di Geofisica e Vulcanologia (INGV), Rome called *ionospheric scintillations arctic campaign coordinated observation* (ISACCO) [39.55] and the north-south scintillation monitoring chain of DLR [39.43]. Such networks allow monitoring of regional scintillation activity or tracking scintillation-associated perturbations propagating from high towards low latitudes or vice versa.

Over South America the LISN multisensoral monitoring network [39.56] contributes essentially to the understanding of the complex physical background of

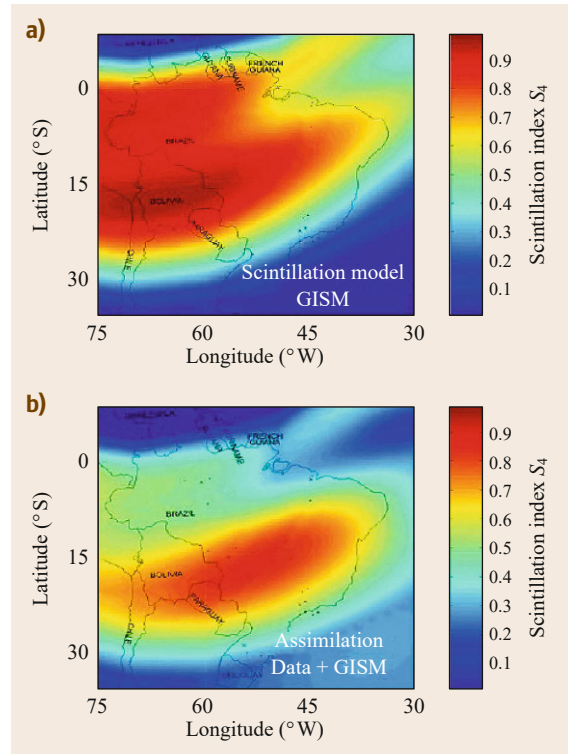


Fig. 39.16 (a) S_4 scintillation map generated by the GISM model for 11 January 2002, at 00:30 UT, solar radio flux $F_{10.7} = 150$, background electron density model is NeQuick. (b) S_4 scintillation map after assimilating measured scintillation data from six Brazilian GPS stations into the GISM background model. More details in [39.52]

the generation and propagation mechanism of ionospheric scintillations. In practical applications single-station information is helpful but usually not sufficient. Regional and/or local GNSS scintillation monitoring networks allow for computation of scintillation maps covering the area in view. Although ionospheric irregularities causing radio scintillations are characterized by small scales, smoothed scintillation maps can provide more reliable information on the current state than climatologic models can do. This may help users at sites where practically no measurements of scintillation activity are available. Thus, actual scintillation measurements assimilated into a climatological scintillation model such as the WBMOD [39.57] and GISM [39.58] improve the model by introducing current geophysical as well as space weather conditions. These conditions have much larger temporal and spatial scales than the irregularities justifying the procedure in a certain sense.

Utilizing GISM as the background model for computing the scintillation activity over Brazil for 10–

18 January 2002, scintillation data obtained from the Brazilian National Institute for Space Research (INPE) were merged into the model [39.52]. Assimilating these actual scintillation data into the GISM model, the obtained scintillation map as seen in Fig. 39.16 is well

39.5 Space Weather

According to the definition given in the national space weather plan of the USA in 1996, space weather refers to the conditions on the Sun and in the solar wind, magnetosphere, ionosphere and thermosphere that can influence the performance and reliability of space-borne and ground-based technological systems and can endanger human life or health. The ionosphere is an integral element of space weather, closely linked to the intensity and the energetic spectrum of electromagnetic and corpuscular radiation emitted from the Sun. The utilization of ground- and space-based GNSS signals is attractive for monitoring space-weather-driven effects in the ionosphere in a twofold manner. Firstly, space-weather-initiated processes can effectively be investigated by probing the ionosphere with high temporal and spatial resolution. Secondly, monitoring results help to develop correction models, mitigation techniques and *ionospheric threat models* to further reduce the ionospheric impact in numerous GNSS applications. The latter point is of interest since requirements on accuracy, spatial resolution, integrity, and continuity of GNSS are permanently growing. In parallel, the robustness of

adapted to the real conditions. The more measurements are available, the higher is the spatial resolution of the maps. The assimilation procedure is similar to that used for routine generation of TEC maps in DLR since 1995 [39.5, 6].

ionospheric measurements, their temporal and spatial resolution and accuracy permanently improve due to the rapidly growing number of ground- and space-based radio links.

As pointed out in Sect. 6.3, the ionospheric ionization is mainly controlled by solar radiation at wavelengths of $< 130\text{ nm}$ and energetic particles originating from the solar wind. Principally, there is a close correlation between TEC and the solar radiation. Whereas photoionization acts immediately, the total ionization follows the solar cycle and solar irradiation changes [39.59] with a delay of 1–2 d [39.60].

To distinguish from climatological effects, in this chapter we focus on space weather effects characterized by timescales of less than about 10 d. Timescales of a few days are typical for the duration of ionospheric storms. Shorter radiation events in the range of minutes are related to solar radiation bursts known as solar flares.

39.5.1 Direct Impact of Solar Radiation and Energetic Particles

If the solar emission spectrum contains a strong enhancement of the ionizing extreme ultraviolet (EUV), the ionospheric range error of GNSS may increase by several meters within a minute. Associated rapid phase changes may lead to problems at receiver level, i.e., seriously degrade positioning and monitoring capabilities. During solar flares the Sun emits electromagnetic waves at a broad frequency spectrum from Gamma via X-rays down to radio waves. This may lead to a sudden increase of TEC (SITEC), as known for many years [39.61, 62].

As shown in Fig. 39.17 (bottom panel) STEC may rapidly jump by 20 TECU or even more during a SITEC event and therefore might reduce accuracy and reliability of GNSS applications to an intolerable extend.

The number of available GPS measurements dropped down from 30 to only 7. Afraimovich et al. [39.62] have discussed the response of global GNSS measurements to faint and bright solar flares. These data can effectively be used to detect in particular solar flares with a strong EUV component in the spectrum [39.63].

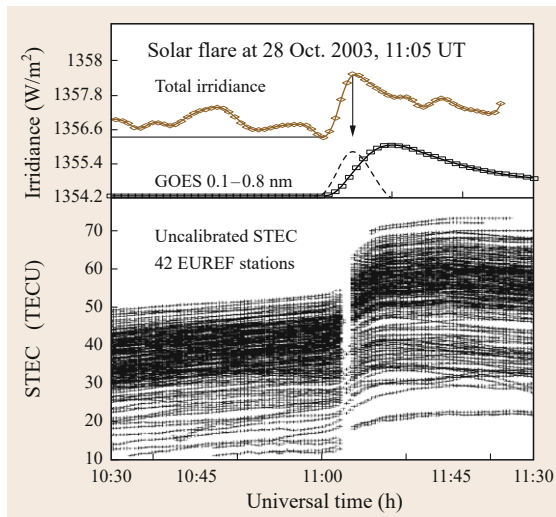


Fig. 39.17 Uncalibrated TEC response of the solar flare on 28 October 2003 at 11:05 UT. Enhancement of the total solar irradiance by 267 ppm caused a TEC jump at all GPS measurements over Europe (range error up to about 3.5 m)

Besides energetic ionizing radiation, also wideband radio waves may be emitted, called radio bursts. On 6 December 2006 the radio burst intensity at GPS frequencies at $L1 = 1575.42$ and $L2 = 1227.60$ MHz was extremely high, causing severe interference problems with GPS measurements at the sunlit side of the Earth [39.64].

SITEC measurements were performed many years ago mainly based on Faraday rotation measurements at linearly polarized beacon signals from geostationary satellites such as ATS 6. Nowadays GNSS measurements are well suited to measure flare-related ionization events. Coinciding with the approach to the 24th solar cycle expected maximum, the interest for a better observation capability of solar flare events, considered as space weather precursors, has grown significantly. In particular, better accuracy and temporal resolution of the changes within the flux of photons are becoming important to get a better understanding of the Sun-Earth relationships [39.65]. GNSS measurements may help to estimate rapid EUV photon flux increases during solar flares [39.63].

Besides electromagnetic radiation, corpuscular radiation of solar origin may also noticeably increase the ionospheric ionization level. The associated TEC increase is measurable both by ground- and space-based dual-frequency GNSS measurements. Since precipitating electrons of magnetospheric origin ionize in particular the bottomside ionosphere, IRO retrievals are well suited to detect associated enhancements of electron density. Thus, numerous IRO datasets of CHAMP and COSMIC/Formosat-3 datasets have been systematically screened to select those profiles that indicate higher ionization at E-layer heights than at F2 layer heights [39.66]. This E-layer-dominated ionosphere (ELDI) is a clear indication of space-weather-related particle precipitation at high latitudes (Fig. 39.18).

As Fig. 39.18b shows, the selected ELDI profiles are well distributed around the auroral oval where fascinating polar lights can be observed. The shape of this Arctic oval has been described by an ellipse indicated by a yellow line in Fig. 39.18b [39.66].

During ionospheric storms the particle precipitation is essentially enhanced, often associated with disruptions of GNSS measurements and services like EGNOS at high latitudes [39.43].

39.5.2 Ionospheric Perturbations and Associated Effects

Ionospheric storms are space-weather-induced large-scale disturbances of the ionospheric structure and dynamics. Due to the strong electrodynamic coupling with the magnetosphere they are closely related to ge-

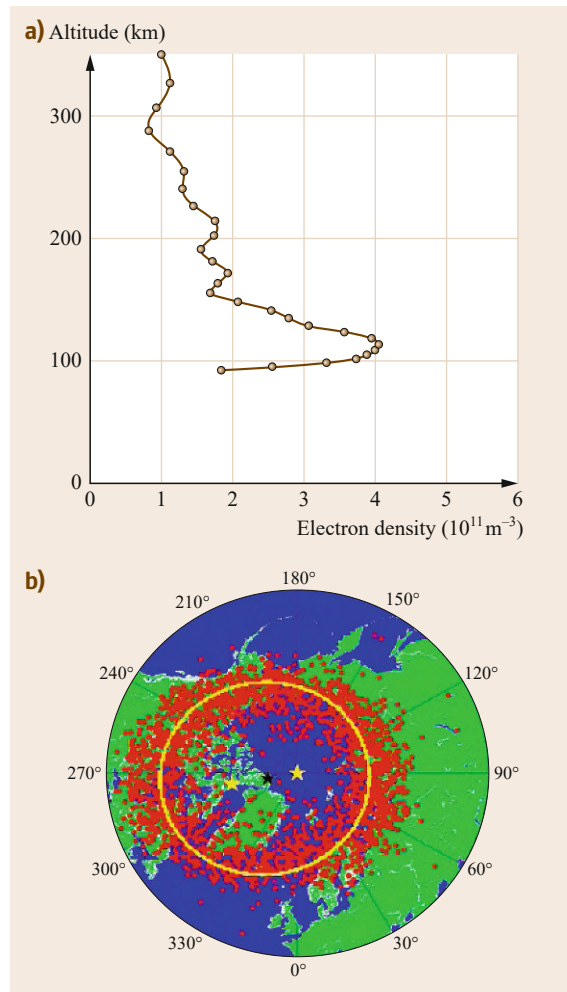


Fig. 39.18 (a) Selected electron density profile retrieved from CHAMP IRO measurements over the South Pole on 29 October 2003 to demonstrate a typical ELDI profile. (b) Ellipse fit to the distribution of COSMIC/Formosat-3 profiles obtained over the Northern hemisphere in January and February 2007 satisfying the ELDI condition. The yellow stars mark the focal points of the ellipse, the black star marks the center of a less accurate circle fit (further details in [39.66])

omagnetic storms. Therefore, ionospheric storms are mostly characterized by geomagnetic indices. Nevertheless, due to the complexity of ionospheric storms, geomagnetic indices fail in describing the storm behavior of the ionospheric plasma. To overcome this problem, attempts are made to characterize ionospheric storms by more specific ionospheric indices [39.67].

Due to the strong coupling with the magnetosphere and the solar wind, enhanced space weather impact is expected in particular at the high-latitude ionosphere

where the geomagnetic field lines come down to the Earth. During severe space weather events a huge amount of solar wind energy couples into the thermosphere/ionosphere/magnetosphere systems thus generating large perturbations in the high-latitude ionosphere and thermosphere. These perturbations are characterized by significant changes in plasma density, composition and temperature accompanied by large-scale plasma transport processes [39.66, 68–71].

A severe ionospheric storm was globally observed at the end of October 2003, called the Halloween storm. The storm was initiated by a huge solar flare of class X17 on 28 October (Fig. 39.17) followed by two severe coronal mass ejections (CMEs) on subsequent days. Whereas there was an immediate TEC response on the flare by more than 10 TECU, persistent large-scale perturbations were observed later when the CMEs reached the Earth on 29 and 30 October 2003. Electromagnetic coupling of the CME plasma cloud with the Earth's magnetosphere causes a complex system of electric fields and currents in the magnetosphere/ionosphere systems. Thus, the magnetospheric dawn-dusk electric field maps down to ionospheric heights along geomagnetic field lines and drives the day-side plasma to the night-side across the poles, creating a so-called tongue of ionization. This is nicely seen in the polar TEC distribution on 29 October at 06:00 and 08:00 UT in Fig. 39.19.

Additionally, huge electric currents in the order of one million ampere are generated in the auroral E-layer ionosphere. The resulting Joule heating of the thermosphere results in numerous effects, such as an expansion of the thermosphere, generation of neutral winds and composition changes. So the ionosphere is heavily disturbed over Europe causing TEC enhance-

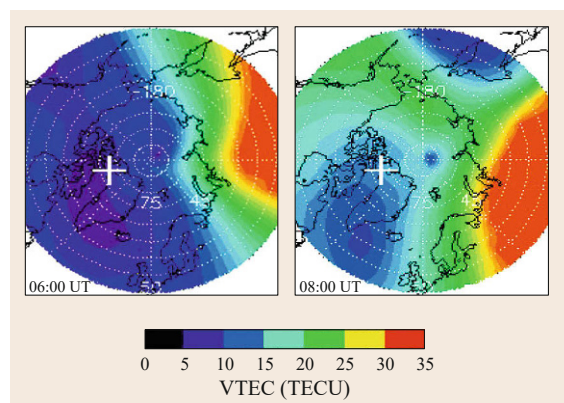


Fig. 39.19 Formation of the tongue of ionization in polar TEC on 29 October 2003 at 06:00 and 08:00 UT. TEC has been derived from ground-based GPS measurements of the IGS station network in DLR (after [39.6])

ments of about 200% at high latitudes. Upwelling of the thermosphere and associated winds initiate transport processes towards lower latitudes, for example large-scale traveling disturbances (LSTIDs) [39.72], which are indicated in Fig. 39.20 by a perturbation pattern derived from ground-based GNSS data.

The related perturbation pattern derived from wavelet analysis is quite useful for studying general features of ionospheric storm generation and propagation [39.73].

The storm pattern seen in Fig. 39.20 indicates quite different plasma transport processes, such as the instantaneous uplifting of plasma throughout all latitudes shown here at about 6:30 UT as a simultaneous enhancement of TEC along all latitudes. This is due to the immediate action of the convection electric field before the ring current has been developed. From 58° N towards the north pole we see the trace of the tongue of ionization seen also in Fig. 39.19. Subsequently a number of perturbation traces are directed towards southern Europe. The slopes indicate velocities in the order of 600 m/s typically for a large-scale perturbation pattern [39.73]. In the afternoon around 15:00 UT the equatorward motion of the mid-latitude trough begins, which separates polar patches in the north from more regular transport processes in the south. The velocity of the southward trough motion is in the order of about 50 m/s.

In the further course of the storm on 29 October 2003 strong plasma uplifting can be observed at the Southern hemisphere at 20:00 UT, i. e., around noon at the right side of corresponding electron density distribution of topside ionosphere/plasmasphere presented in Fig. 39.21.

Strong ionization enhancement is shown also in the equatorial region indicating the action of a very strong eastward-directed electric field, which is consistent with the southward motion of the mid-latitude trough in the evening hours at the European sector, noted before.

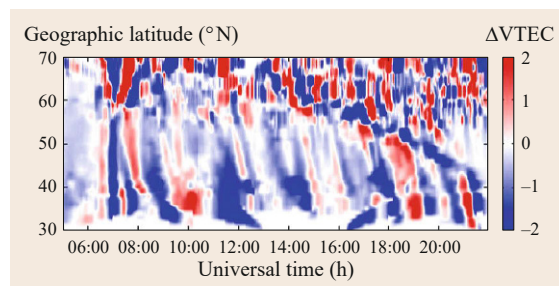


Fig. 39.20 TID storm pattern in TEC observed during the Halloween storm on 29 October 2003 along the 12° E meridian across Europe

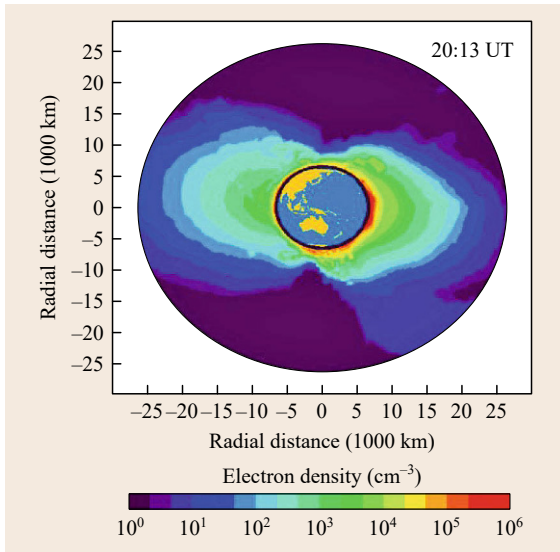


Fig. 39.21 Topside reconstruction of the electron density distribution in the CHAMP orbit plane on 29 October 2003 at 20:13 UT, (after [39.74])

It is worth mentioning that also the corresponding IRO measurements on board CHAMP are consistent with this interpretation. They provide even more insight into the dynamics of the ionosphere during the severe Halloween storm [39.74]. Summarizing, it can be stated that ground- and space-based dual-frequency GNSS measurements provide a powerful tool for probing the ionosphere, in particular under perturbed conditions when other techniques such as vertical sounding may fail.

Considering the GNSS user side, this Halloween storm has demonstrated the reality of ionospheric threats on global scale, for example when WAAS over the US failed for several hours [39.75].

39.5.3 Prediction of Space Weather Phenomena

Space weather studies reported in the previous sections shall provide an improved understanding of the physics behind space weather phenomena in order to better forecast space-weather-related effects and their impact on technical systems. Principally, the GNSS user community is interested in warnings and forecasts of the ionospheric behavior in particular during perturbations. This is a challenging task that cannot be managed on the basis of empirical models such as IRI or NeQuick, which are principally climatological. Therefore, physics-based models are needed that are permanently updated by current observation data. Data-driven physics-based models such as GAIM developed at Utah State University [39.33] or the CTIPE model operated at the Space Weather Prediction Center (SWPC) in Boulder, Colorado, [39.76] have the principal capability to forecast ionospheric behavior in a sufficient way. These complex models suffer from the lack of specific input data needed to get unambiguous solutions of the system of commonly used physical equations, such as the continuity equation, equations of motion and energy equations. Taking into account that the data situation with respect to ground- and space-based GNSS and complementary ionospheric data sources is permanently improving, the develop-

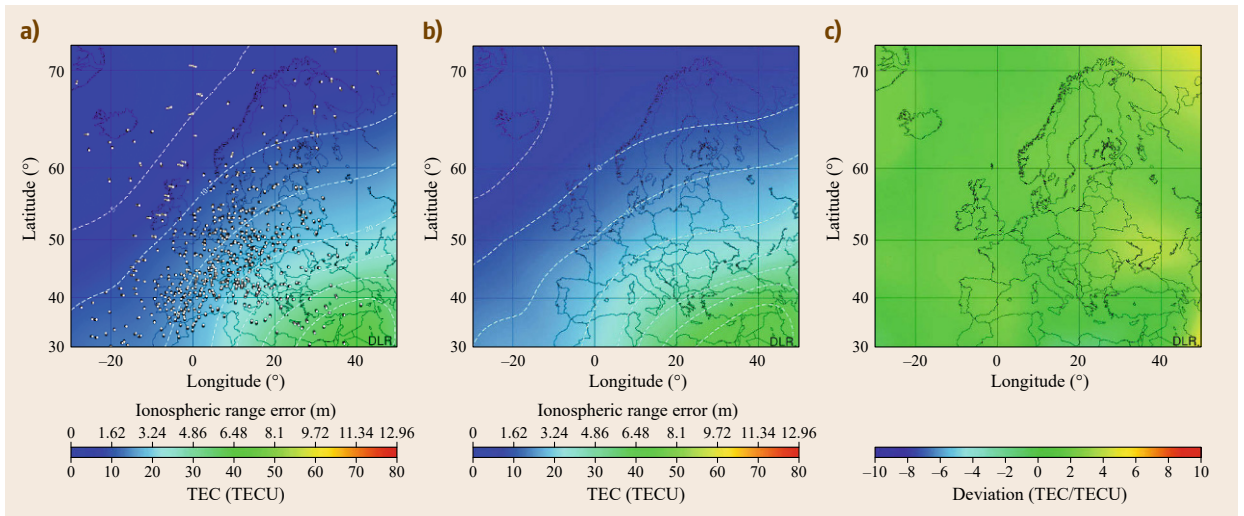


Fig. 39.22a–c VTEC map over Europe on 17 June 2012, 06:00 UT as provided via SWACI (a), 1 h VTEC forecast for Europe (b), quality of forecast released at 05:00 UT for 06:00 UT (c)

ment of data-driven physics-based models is the only way to get reliable forecasts of the ionospheric behavior under perturbed conditions in the future. In the meantime pragmatic solutions based on near-real-time measurements of the current state of the ionosphere and its drivers like solar radiation and solar wind are still justified. Accordingly, there are still some international efforts to derive forecasts from a combined use of actual data and empirical models or neural networks. To learn more about physics-based models, comparison of results obtained from physics-based models and GNSS-derived TEC maps or 3-D reconstructions of the electron density are very helpful [39.76].

As an example, TEC map forecasts of 1 h ahead are made routinely via the SWACI service of DLR [39.8], which is based on the current ionospheric behavior and a background model (Fig. 39.22). To estimate the quality of the previous forecast, it is checked with real data 1 h later when the forecast time is reached. The corresponding difference plot is available for users allowing immediate estimation of the quality of the forecast. Mean forecast errors of the SWACI service are usually less than 10% of the original values.

To further improve the current forecast quality, the development of an empirical storm model controlled by solar wind parameters received from the ACE satellite is envisaged.

39.6 Coupling with Lower Geospheres

Although space weather impacts on the ionospheric plasma clearly dominate, ionospheric plasma may be affected by processes propagating from lower geospheres such as the lower atmosphere and even the litho- or hydrosphere. Such processes are briefly discussed in the two subsequent sections.

39.6.1 Atmospheric Signatures

The Earth's atmosphere is periodically excited by solar radiation. Taking into account the rotation of the Earth, the permanently changing inclination and the gravity impact of the Moon, various wavelike processes are excited in the atmosphere on planetary and regional scale. In addition to these well-defined regular processes, the solar radiation as the main driving force shows quasiperiodical variations, for example associated with the mean solar rotation period of 27 d and the solar cycle period of about 11 y.

Here we consider a kind of *second-order effect* on the balance of ionospheric plasma, namely the impact of wave phenomena that have been excited in the lower atmosphere due to the above-mentioned natural circumstances.

The propagation of hydrodynamic motions such as gravity waves, tides and planetary waves has been discussed already for more than five decades. Assuming conservation of energy flux of an upward propagating wave, their amplitude increases rapidly with height due to the exponential decrease of neutral gas density. This is the main reason why waves that are small at low heights may reach amplitudes that are sufficiently large to be measurable at ionospheric altitudes. Having a more detailed view, one has to consider in practice characteristics of the atmospheric filter function for the

vertical propagation of atmospheric waves like planetary waves (PWs).

Planetary waves depend in particular on the varying vertical component of rotation with geographic latitude. These waves manifest in large-scale variations in neutral wind, density and pressure propagating zonally and vertically from troposphere-stratosphere regions towards middle atmosphere and lower thermosphere regions. Their oscillation periods are in the order of 2–20 d with main periods at 2, 5, 10 and 16 d [39.77].

Another type of wave occurring in the atmosphere are called gravity waves. In the case of an initial perturbation, for example heating or vertical displacement in a stratified atmosphere, the restoring force is gravity.

Both wave types distinguish further in their capabilities to propagate upward. Whereas planetary waves cannot penetrate the turbopause region at about 110 km height due to physical reasons, gravity waves are able to propagate through it. Although there is an upper boundary for upward propagating planetary waves, planetary-type oscillations of the ionospheric plasma up to F2 layer heights have been observed by airglow and radar measurements and ionospheric sounding [39.78].

Since PWs cannot overcome the 110 km barrier, some supporting mechanism must exist to convert the wave energy of PWs into another atmospheric process that becomes visible in higher ionospheric layers. Observational evidence has shown that the ionospheric system of currents in the E-layer or dynamo region at around 100 km height might directly be modulated by planetary waves [39.78]. The observation of typical PW oscillation periods in the equatorial electrojet indicates a significant influence of planetary waves on plasma parameters of the equatorial ionosphere. The associated electric field could also be responsible for observed nighttime F-

layer height oscillations at PW periodicities. Generally speaking, there are many possible scenarios to explain planetary wave signatures in the ionosphere [39.77]. Considering the episodic character of PW and the difficulty of distinguishing their ionospheric signatures from solar or more complex space weather forcing [39.79], TEC measurements will substantially help to explore the physics behind the related atmosphere-ionosphere coupling, in particular if dense GNSS networks are available. To give an example, Fig. 39.23 shows the existence of PW-type oscillations in TEC [39.80], which may contribute to the range error in GNSS as pointed out earlier. Although a number of observations provide experimental evidence for the existence of PW-type oscillations even in the F2 layer ionosphere, more comprehensive studies are required to get deeper insight into the vertical coupling processes of planetary waves in the mesosphere, thermosphere and ionosphere.

Besides planetary waves or space-weather-driven large-scale storm patterns, medium-scale traveling disturbances (MSTIDs) can also be monitored by ground- [39.81] and space-based [39.82] TEC measurements.

MSTIDs, which may affect precise positioning, can be considered as ionospheric signatures of atmospheric gravity waves (AGWs) reaching amplitudes of up to few TECUs. Wave periods are in the order of several minutes up to about 1 h at horizontal wavelengths of a few hundred kilometers and velocities reaching from 50–300 m/s [39.81]. The observed AGWs may be associated with meteorological phenomena such as thunderstorms, solar eclipses or the solar terminator. GNSS networks are well suited to analyze amplitude, horizontal direction, speed, frequency and wavelength of such waves.

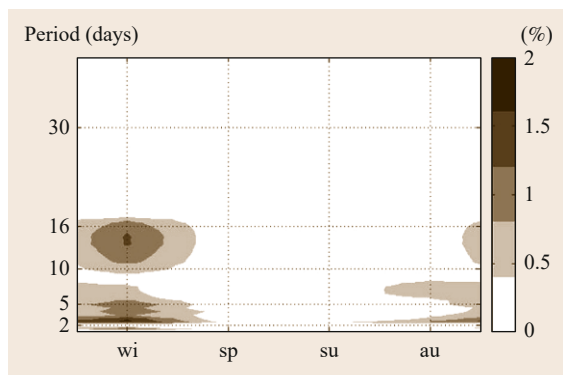


Fig. 39.23 Sample for the occurrence of typical planetary wave periods (westward directed waves of wave no.1) in the total electron content of the ionosphere representative for three months in the four seasons winter (wi), spring (sp), summer (su) and autumn (au). More details in [39.80]

Thus, the solar terminator (ST) may excite two types of AGWs; a long-period AGW (≈ 60 min) with amplitudes in the order of 0.5–1 TECU and a short-period one ≈ 15 min) one with amplitudes in the order of 0.05–0.1 TECU [39.83].

Due to the presence of wavelike perturbations in the ionosphere, positioning errors in networks for real-time kinematic (RTK) positioning (Chap. 26) can reach about 25 cm for a 25 km baseline [39.85]. Since MSTIDs are moving structures such as LSTIDs shown in Fig. 39.20, their effect on positioning varies with the baseline orientation. Thus, if a given baseline is oriented perpendicular to the MSTID propagation direction, the related positioning error will be smaller than that observed for a baseline oriented parallel to the propagation direction. GNSS-based ionospheric monitoring contributes to the mitigation of this problem by investigating the physical processes causing the temporal and spatial characteristics of MSTIDs in the application region. Amplitudes, typical speeds, and direction of propagation were, for example, analyzed for numerous datasets in relation to geophysical conditions in [39.84]. To give an example, velocities and propagation direction of MSTIDs derived from GPS measurements in California between 2004 and 2011 are shown in Fig. 39.24.

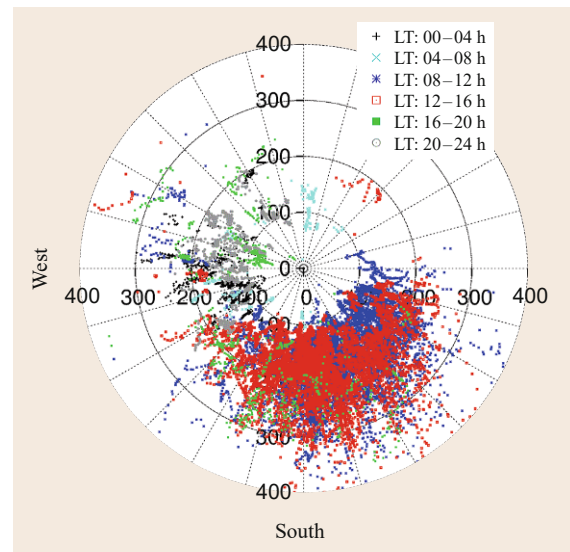


Fig. 39.24 Polar plot of MSTID velocities (m/s) as a function of azimuth representative for winter season in California averaged over years 2004–2011. Local time dependence is parameterized by colors: *black* for LT 00–04 h, *light blue* for LT 04–08 h, *dark blue* for LT 08–12 h, *red* for LT 12–16 h, *green* for LT 16–20 h and *gray* for LT 20–24 h (after [39.84], courtesy of John Wiley)

Taking into account these relationships, an empirical model has been developed in [39.84] that describes climatological features of MSTIDs as a function of geophysical and solar activity conditions. Besides characterizing wavelike processes in the ionosphere whose origin is still a matter of research, the model helps to estimate ionospheric threats in RTK positioning.

39.6.2 Earthquake Signatures

Besides atmospheric perturbations, also strong perturbations in the Earth's litho- and hydrosphere may generate signatures in the ionospheric plasma density via atmospheric-ionospheric coupling processes. Due to the sensitivity of the differential carrier-phase to ionospheric ionization changes, ground- and space-based dual-frequency GPS measurements offer a unique opportunity for detecting earthquake and tsunami signatures in the ionosphere. Vertical displacements of the Earth's surface may excite pressure or acoustic waves that propagate upward. Despite their small amplitude also tsunami waves may excite the above-lying atmosphere by generating gravity waves that propagate obliquely upward.

Since the atmospheric density decreases almost exponentially with altitude, energy conservation implies that the wave amplitude increases exponentially as mentioned already in the previous section. Hence, the amplification may reach a factor of 10^4 – 10^6 in a limited frequency range with periods in the order of 2–6 min. GPS-based detection of earthquake signatures in the ionosphere was first reported by *Calais and Minster* [39.86]. While analyzing the Denali earthquake on 3 November 2002, further evidence was brought that differential GPS phases can measure ionospheric plasma changes induced by upward-propagating atmospheric acoustic waves [39.87, 88]. Earthquake-associated acoustic waves need approximately 10 min to reach the F-layer of the ionosphere, where the close coupling between the neutral atmosphere and ionized plasma results in a wavelike variation of the electron density. Differential GPS measurements made at Sampali, Indonesia (3.62° N; 98.71° E) during the severe earthquake on 26 December 2006 in Indonesia indicate a significant signal about 10 min after the earthquake shock (Fig. 39.25). About 2 h after the event the ionosphere was completely recovered [39.88].

Besides earthquake signatures also associated oceanographic tsunami signatures can be seen in the ionosphere plasma [39.89, 90]. They may excite gravity waves in the troposphere, which propagate upward toward F-layer heights. In contrast to the acoustic waves these waves have longer periods in the typical range of

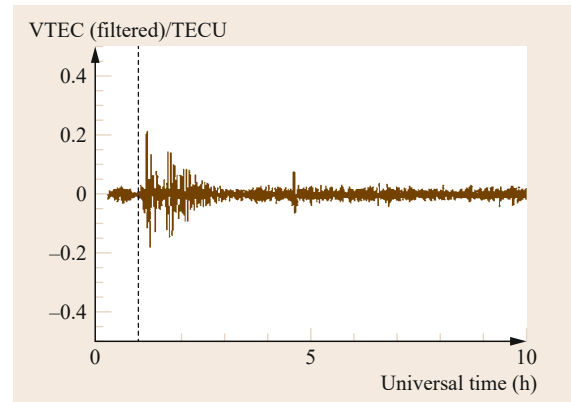


Fig. 39.25 Band-pass filtered VTEC (2.5–10 min) data from GNSS station Sampali at 3.62° N; 98.71° E during the severe Sumatra earthquake on 26 December 2004 in Indonesia. Onset of the earthquake is indicated by a dashed line at 00:58:53 UT (after [39.88])

10–30 min. The travel time to reach ionospheric heights is about 2 h.

Although various earthquake parameters such as the depth influence the detectability of related signatures in the ionosphere, it can be stated that earthquakes of magnitude 6 and higher are detectable by GNSS measurements in general. Related GNSS studies help to explore the coupling mechanisms between lithosphere, hydrosphere and ionosphere including different atmospheric layers. A big advantage of the GNSS technique used for tsunami detection is the large observation range reaching up to more than 1000 km from the ground station over the ocean. For estimating the statistical significance of the measurement results, the availability of dense and widespread networks of permanent GNSS receivers is required.

Instead of measuring after effects it would be of great social importance to detect precursor effects of earthquakes a few hours or days in advance. There exist several publications on this matter suggesting diverse mechanisms for how precursor effects could act and trying to provide evidence for the existence of precursor effects by various types of observations, among them GNSS observations [39.91].

Although pre-earthquake TEC anomalies were found in some studies [39.92] there was no statistically significant correlation observed between TEC anomalies and the occurrence of earthquakes in southern California for the period 2003–2004 [39.93]. Since the key question of whether precursor effects in the ionosphere exist is still open, the search for ionospheric precursor effects continues. GNSSs provide a powerful tool for monitoring associated ionospheric effects and might serve as part of an earthquake warning system.

39.7 Information and Data Services

Ionospheric disturbances can adversely affect ground- and space-based systems and operations, including over-the-horizon radars, HF communications, PNT services and remote sensing radars. Hence, operators of these systems are interested in the newest information on current space weather conditions.

The International Space Environment Service (ISES) is a collaborative network of space weather service-providing organizations around the globe. The task of ISES is to improve, to coordinate, and to deliver operational space weather services. ISES is organized and operated for the benefit of the international space weather user community. The service currently includes 14 regional warning centers, four associate warning centers, and one collaborative expert center. ISES is a network member of the International Council for Science World Data System (ICSU-WDS) and collaborates with the World Meteorological Organization (WMO) and other international organizations.

A few of the ISES centers provide GNSS-derived near-real-time regional TEC maps: INPE (Brazil), NICT (Japan), NOAA (USA), and the Radio and Space Weather Services, Bureau of Meteorology (Australia).

Near-real-time (delay ≤ 30 min) global TEC maps are generated and provided by DLR (Germany), UPC (Spain), International GNSS Service (IGS), JPL (USA), and Utah State University Space weather Center (USA).

Space-based GNSS data and related retrievals, for example GPS radio-occultation data, from LEO satellites such as COSMIC/Formosat-3 and GRACE are provided by DLR (Germany), Taiwan Analysis Center for COSMIC (Taiwan) and UCAR (USA).

In order to permanently monitor the ionospheric state and in particular to detect and trace space weather effects, powerful GNSS-based monitoring services have been established in European countries.

Within their Space Situational Awareness (SSA) program the European Space Agency (ESA) is establishing five expert service centers related to solar weather, space radiation, ionosphere, geomagnetic field and heliosphere. The SSA preparatory program was started in 2009. The SSA space weather segment is planned to be further developed in upcoming years by exploiting the European expertise in the space weather area under the coordination of ESA.

The WMO supports international coordination of space weather activities and services. In May 2010, WMO established the interprogram coordination team on space weather (ICTSW) with a mandate to support space weather observation, data exchange, product and services delivery, and operational applications. ICTSW involves experts from numerous countries and international organizations.

Acknowledgments. The author would like to express his gratitude to colleagues from DLR's Institute of Communications and Navigation and Earth Observation Center for the close collaboration on many projects, e.g., related to satellite missions CHAMP and GRACE and the space weather project SWACI. He would like to thank the international geodetic community, in particular the International GNSS Service, for providing free access to high quality GNSS data for more than two decades.

References

- 39.1 B.D. Wilson, A.J. Mannucci: Instrumental biases in ionospheric measurements derived from GPS data, *Proc. Inst. Nav.* **93**(2), 1343–1351 (1993)
- 39.2 E. Sardón, A. Rius, N. Zarraoa: Estimation of the receiver differential biases and the ionospheric total electron content from global positioning system observations, *Radio Sci.* **29**, 577–586 (1994)
- 39.3 L. Ciraolo, P. Spalla, P. Beni: An analysis of consistency of TEC evaluated using pseudo-range GPS observations, *Proc. Int. Beacon Satell. Symp.*, Aberystwyth, ed. by L. Kersley (Univ. Wales, Aberystwyth 1994) pp. 21–24
- 39.4 M. Hernández-Pajares, J.M. Juan, J. Sanz: New approaches in global ionospheric determination using ground GPS data, *J. Atmos. Sol.-Terr. Phys.* **61**(16), 1237–1247 (1999)
- 39.5 N. Jakowski: *TEC Monitoring by Using Satellite Positioning Systems*, ed. by H. Kohl, R. Rüster, K. Schlegel (European Geophysical Society, Katlenburg-Lindau 1996) pp. 371–390
- 39.6 N. Jakowski, C. Mayer, M.M. Hoque, V. Wilken: Total electron content models and their use in ionosphere monitoring, *Radio Sci.* **46**(RS0D18), 1–11 (2011)
- 39.7 N. Jakowski, M.M. Hoque, C. Mayer: A new global TEC model for estimating transionospheric radio wave propagation errors, *J. Geod.* **85**(12), 965–974 (2011)
- 39.8 N. Jakowski, C. Mayer, K.-D. Missling, H. Barkmann, C. Borries, H. Maas, T. Noack, M. Tegler, V. Wilken: Products and services provided by the Space Weather Application Center – Ionosphere (SWACI), *Proc. Space Weather Work.*, Boulder (NOAA, Washington DC 2010) pp. 1–23
- 39.9 J.M. Dow, R.E. Neilan, C. Rizos: The international GNSS service in a changing landscape of global

- navigation satellite systems, *J. Geod.* **83**(3–4), 191–198 (2009)
- 39.10 M. Luo, S. Pullen, H. Dennis, J. Konno, G. Xie, T. Walter, P. Enge, S. Datta-Barua, T. Dehel: LAAS ionosphere spatial gradient threat model and impact of LGF and airborne monitoring, *Proc. ION GPS* (2003) pp. 2255–2274
- 39.11 C. Mayer, B. Belabbas, N. Jakowski, M. Meurer, W. Dunkel: Ionosphere Threat Space Model Assessment for GBAS, *Proc. ION GNSS, Savannah* (2009) pp. 1091–1099
- 39.12 G.A. Hajj, L.J. Romans: Ionospheric electron density profiles obtained with the Global Positioning System: Results from the GPS/MET experiment, *Radio Sci.* **33**(1), 175–190 (1998)
- 39.13 N. Jakowski, A. Wehrenpfennig, S. Heise, C. Reigber, H. Lühr, L. Grunwaldt, T.K. Meehan: GPS radio occultation measurements of the ionosphere from CHAMP: Early results, *Geophys. Res. Lett.* **29**(10), 95–1–95–4 (2002)
- 39.14 N. Jakowski: Ionospheric GPS radio occultation measurements on board CHAMP, *GPS Solutions* **9**(2), 88–95 (2005)
- 39.15 S. Heise, N. Jakowski, A. Wehrenpfennig, C. Reigber, H. Lühr: Sounding of the topside ionosphere/plasmasphere based on GPS measurements from CHAMP: Initial results, *Geophys. Res. Lett.* **29**(14), 44.1–44.4 (2002)
- 39.16 C. Rocken, Y.-H. Kuo, W. Schreiner, D. Hunt, S. Sokolovskiy, C.M. Cormick: COSMIC system description, *Terr. Atmos. Ocean. Sci. (Special issue)* **11**(1), 21–52 (2000)
- 39.17 T.P. Yunck, F. Lindal, C.H. Liu: The role of GPS in precise Earth observation, *Proc. IEEE PLANS, Orlando* (1988) pp. 251–258, doi: 10.110g/PLANS.1988.195491
- 39.18 G.A. Fjeldbo, J. Kliore, V.R. Eshleman: The neutral atmosphere of Venus as studied with the Mariner V radio occultation experiments, *Astron. J.* **76**(2), 123–140 (1971)
- 39.19 J. Wickert, C. Reigber, G. Beyerle, R. König, C. Marquardt, T. Schmidt, L. Grunwaldt, R. Galas, T.K. Meehan, W.G. Melbourne, K. Hocke: Atmosphere sounding by GPS radio occultation: First results from CHAMP, *Geophys. Res. Lett.* **28**(17), 3263–3266 (2001)
- 39.20 M.M. Hoque, N. Jakowski: Higher order ionospheric propagation effects on GPS radio occultation signals, *Adv. Space Res.* **46**(2), 162–173 (2010)
- 39.21 M.M. Hoque, N. Jakowski: Ionospheric bending correction for GNSS radio occultation signals, *Radio Sci.* **46**(RS0D06), 1–9 (2011)
- 39.22 M.M. Hoque, N. Jakowski: A new global empirical NmF2 model for operational use in radio systems, *Radio Sci.* **46**(RS6015), 1–13 (2011)
- 39.23 M.M. Hoque, N. Jakowski: A new global model for the ionospheric F2 peak height for radio wave propagation, *Ann. Geophys.* **30**(5), 797–809 (2012)
- 39.24 K. Davies: *Ionospheric Radio* (Peter Peregrinus, London 1990)
- 39.25 D.B. Mularew: Alouette-ISIS radio wave studies of the cleft, the auroral zone, and the main trough and of their associated irregularities, *Radio Sci.* **18**(6), 1140–1150 (1983)
- 39.26 P.L. Timleck, G.L. Nelms: Electron densities less than 100 electron cm⁻³ in the topside ionosphere, *Proc. IEEE* **57**, 1164–1171 (1969)
- 39.27 R.E. Daniell, L.D. Brown, D.N. Anderson, M.W. Fox, P.H. Doherty, D.T. Decker, J.J. Sojka, R.W. Schunk: Parameterized ionospheric model: A global ionospheric parameterization based on first principles models, *Radio Sci.* **30**(5), 1499–1510 (1995)
- 39.28 J.R. Austen, S.J. Franke, C.H. Liu: Ionospheric imaging using computerized tomography, *Radio Sci.* **23**(3), 299–307 (1988)
- 39.29 L. Kersley, S.E. Pryse: Development of experimental ionospheric tomography, *Int. J. Imaging Syst. Technol.* **5**(2), 141–147 (1994)
- 39.30 L. Kersley, S.E. Pryse, M.H. Denton, G. Bust, E. Fremouw, J. Secan, N. Jakowski, G.J. Bailey: Radio tomographic imaging of the northern high-latitude ionosphere on a wide geographic scale, *Radio Sci.* **40**(RS5003), 1–9 (2005)
- 39.31 G.S. Bust, D. Coco, J.J. Makela: Combined ionospheric campaign 1: Ionospheric tomography and GPS total electron count (TEC) depletions, *Geophys. Res. Lett.* **27**(18), 2849–2852 (2000)
- 39.32 G.S. Bust, T.W. Garner, T.L. Gaussiran: Ionospheric data assimilation three-dimensional (IDA3D): A global, multisensor, electron density specification algorithm, *J. Geophys. Res. Space Phys.* **109**(A11), 1–14 (2004)
- 39.33 R.W. Schunk, L. Scherliess, J.J. Sojka, D.C. Thompson, D.N. Anderson, M. Codrescu, C. Minter, T.J. Fuller-Rowell, R.A. Heelis, M. Hairston, B.M. Howe: Global assimilation of ionospheric measurements (GAIM), *Radio Sci.* **39**(RS1S02), 1–11 (2004)
- 39.34 C.N. Mitchell, P.S.J. Spencer: A three-dimensional time-dependent algorithm for ionospheric imaging using GPS, *Ann. Geophys.* **46**(4), 687–696 (2003)
- 39.35 M.J. Angling, P.S. Cannon: Assimilation of radio occultation measurements into background ionospheric models, *Radio Sci.* **39**(RS1S0), 1–11 (2004)
- 39.36 L. Mandrake, B. Wilson, C. Wang, G. Hajj, A. Manucci, X. Pi: A performance evaluation of the operational jet propulsion laboratory/University of southern California global assimilation ionospheric model (JPL/USC GAIM), *J. Geophys. Res. Space Phys.* **110**(A12306), 1–10 (2005)
- 39.37 G.S. Bust, C.N. Mitchell: History, current state, and future directions of ionospheric imaging, *Rev. Geophys.* **46**(1), 1–23 (2008)
- 39.38 D. Bilitza, B.W. Reinisch: International reference ionosphere 2007: Improvements and new parameters, *Adv. Space Res.* **42**(4), 599–609 (2008)
- 39.39 P.M. Kintner, B.M. Ledvina: The ionosphere, radio navigation, and global navigation satellite systems, *Adv. Space Res.* **35**(5), 788–811 (2005)
- 39.40 J. Aarons, C. Gurguolo, A.S. Rodger: The effects of magnetic storm phases on F layer irregularities below the auroral oval, *Radio Sci.* **23**(3), 309–319 (1988)
- 39.41 S. Basu, E. Kudeki, S. Basu, C.E. Valladares, E.J. Weber, H.P. Zengingonul, S. Bhattacharyya, R. Shee-

- han, J.W. Meriwether, M.A. Biondi, H. Kuenzler, J. Espinoza: Scintillations, plasma drifts, and neutral winds in the equatorial ionosphere after sunset, *J. Geophys. Res. Space Phys.* **101**(A12), 26795–26809 (1996)
- 39.42 L. Alfonsi, L. Spogli, J.R. Tong, G. De-Franceschi, V. Romano, A. Bourdillon, M. Le Huy, C.N. Mitchell: GPS scintillation and TEC gradients at equatorial latitudes in April 2006, *Adv. Space Res.* **47**(10), 1750–1757 (2011)
- 39.43 N. Jakowski, Y. Béniguel, G. De-Franceschi, M. Hernández-Pajares, K.S. Jacobsen, I. Stanislawski, L. Tomasik, R. Warnant, G. Wautelet: Monitoring, tracking and forecasting ionospheric perturbations using GNSS techniques, *J. Space Weather Space Clim.* **2**(A22), 1–14 (2012)
- 39.44 S. Basu, E. MacKenzie, S. Basu: Ionospheric constraints on VHF/UHF communications links during solar maximum and minimum periods, *Radio Sci.* **23**(3), 363–378 (1988)
- 39.45 B. Arbesser-Rastburg, N. Jakowski: Effects on satellite navigation. In: *Space Weather: Physics and Effects*, ed. by V. Bothmer, I.A. Daglis (Springer, Heidelberg 2007) pp. 383–402
- 39.46 J.J. Valette, P. Lassudrie-Duchesne, N. Jakowski, Y. Béniguel, V. Wilken, M. Cueto, A. Bourdillon, C. Pollara-Brevart, P. Yaya, J.P. Adam, R. Fleury: Observations of ionospheric perturbations on GPS signals at 50 Hz, 1 Hz and 0.03 Hz in South America and Indonesia, *Proc. 4th Eur. Space Weather Week*, Brussels (Royal Observatory of Belgium, Brussels 2007)
- 39.47 F.M. Dújanga, P. Baki, J.O. Olwendo, B.F. Twina-masiko: Total electron content of the ionosphere at two stations in East Africa during the 24–25 October 2011 geomagnetic storm, *Adv. Space Res.* **51**(5), 712–721 (2013)
- 39.48 M. Nishioka, A. Saito, T. Tugawa: Occurrence characteristics of plasma bubble derived from global ground-based GPS receiver networks, *J. Geophys. Res. Space Phys.* **113**(A05301), 1–12 (2008)
- 39.49 T.L. Beach, P.M. Kintner: Simultaneous global positioning system observations of equatorial scintillations and total electron content fluctuations, *J. Geophys. Res. Space Phys.* **104**(A10), 22553–22565 (1999)
- 39.50 A. Bhattacharyya, T.L. Beach, S. Basu, P.M. Kintner: Nighttime equatorial ionosphere: GPS scintillations and differential carrier phase fluctuations, *Radio Sci.* **35**(1), 209–224 (2000)
- 39.51 C.S. Carrano, K. Groves: The GPS segment of the AFRL-SCINDA global network and the challenges of real-time TEC estimation in the equatorial ionosphere, *Proc. ION ITM*, Monterey (2006) pp. 1036–1047
- 39.52 Y. Béniguel, J.-P. Adam, N. Jakowski, T. Noack, V. Wilken, J.-J. Valette, M. Cueto, A. Bourdillon, P. Lassudrie-Duchesne, B. Arbesser-Rastburg: Analysis of scintillation recorded during the PRIS measurement campaign, *Radio Sci.* **44**(RS0A3), 1–11 (2009)
- 39.53 R. Prieto-Cerdeira, Y. Béniguel: The MONITOR project: Architecture, data and products, *Proc. Ionos. Eff. Symp.*, Alexandria (2011) pp. 1–6
- 39.54 O. de La Beaujardière: C/NOFS: A mission to forecast scintillations, *J. Atmos. Solar-Terr. Phys.* **66**(17), 1573–1591 (2004)
- 39.55 G. Franceschi, L. Alfonsi, V. Romano: ISACCO: An Italian project to monitor the high latitudes ionosphere by means of GPS receivers, *GPS Solutions* **10**(4), 263–267 (2006)
- 39.56 C.E. Valladares, P.H. Doherty: The low-latitude ionosphere sensor network (LISN), *Proc. ION ITM* (2009) pp. 16–24
- 39.57 J.A. Secan, R.M. Bussey, E.J. Fremouw, S. Basu: High-latitude upgrade to the wideband ionospheric scintillation model, *Radio Sci.* **32**(4), 1567–1574 (1997)
- 39.58 Y. Béniguel, P. Hamel: A global ionosphere scintillation propagation model for equatorial regions, *J. Space Weather Space Clim.* **1**(1), A04 (2011)
- 39.59 J.L. Lean, T.N. Woods: Solar spectral irradiance: Measurements and models. In: *Heliophysics: Evolving Solar Activity and the Climates of Space and Earth*, ed. by C.J. Schrijver, G.L. Siscoe (Cambridge Univ. Press, Cambridge 2010) pp. 269–298
- 39.60 N. Jakowski, B. Fichtelmann, A. Jungstand: Solar activity control of ionospheric and thermospheric processes, *J. Atmos. Terr. Phys.* **53**(11/12), 1125–1130 (1991)
- 39.61 K. Davies: Recent progress in satellite radio beacon studies with particular emphasis on the ATS-6 radio beacon experiment, *Space Sci. Rev.* **25**(4), 357–430 (1980)
- 39.62 E.L. Afraimovich, A.T. Altynsev, V.V. Grechnev, L.A. Leonovich: The response of the ionosphere to faint and bright solar flares as deduced from global GPS network data, *Ann. Geophys.* **45**(1), 31–40 (2002)
- 39.63 A. García-Rigo, M. Hernández-Pajares, J.M. Juan, J. Sanz: Solar flare detection system based on global positioning system data: First results, *Adv. Space Res.* **39**(5), 889–895 (2007)
- 39.64 A.P. Cerruti, P.M. Kintner, D.E. Gary, L.J. Lanzerotti, E.R. de Paula, H.B. Vo: Observed solar radio burst effects on GPS/wide area augmentation system carrier-to-noise ratio, *Space Weather* **4**(10), 1–9 (2006)
- 39.65 T.N. Woods, R. Hock, F. Eparvier, A.R. Jones, P.C. Chamberlin, J.A. Klimchuk, L. Didkovsky, D. Judge, J. Mariska, H. Warren, C.J. Schrijver, D.F. Webb, S. Bailey, W.K. Tobiska: New solar extreme-ultraviolet irradiance observations during flares, *Astrophys. J.* **739**(2), 1–13 (2011)
- 39.66 C. Mayer, N. Jakowski: Enhanced E-layer ionization in the auroral zones observed by radio occultation measurements onboard CHAMP and Formosat-3/COSMIC, *Ann. Geophys.* **27**(3), 1207–1212 (2009)
- 39.67 N. Jakowski, C. Borries, V. Wilken: Introducing a disturbance ionosphere index, *Radio Sci.* **47**(RS0L14), 1–9 (2012)
- 39.68 G.W. Prölss: Ionospheric F-region storms. In: *Handbook of Atmospheric Electrodynamics*, ed. by H. Volland (CRC, Boca Raton 1995) pp. 195–248

- 39.69 C.M. Ho, A.J. Mannucci, U.J. Lindqwister, X. Pi, B.T. Tsurutani: Global ionosphere perturbations monitored by the worldwide GPS network, *Geophys. Res. Lett.* **23**(22), 3219–3222 (1996)
- 39.70 N. Jakowski, S. Schlüter, E. Sardon: Total electron content of the ionosphere during the geomagnetic storm on 10 January 1997, *J. Atmos. Solar-Terr. Phys.* **61**(3/4), 299–307 (1999)
- 39.71 M. Förster, N. Jakowski: Geomagnetic storm effects on the topside ionosphere and plasmasphere: A compact tutorial and new results, *Surv. Geophys.* **21**(1), 47–87 (2000)
- 39.72 T. Tsugawa, A. Saito, Y. Otsuka: A statistical study of large-scale traveling ionospheric disturbances using the GPS network in Japan, *J. Geophys. Res. Space Phys.* **109**, A06302 (2004)
- 39.73 C. Borries, N. Jakowski, V. Wilken: Storm induced large scale TIDs observed in GPS derived TEC, *Ann. Geophys.* **27**(4), 1605–1612 (2009)
- 39.74 N. Jakowski, V. Wilken, C. Mayer: Space weather monitoring by GPS measurements on board CHAMP, *Space Weather* **5**, 1–23 (2007)
- 39.75 A. Komjathy, L. Sparks, A.J. Mannucci, A. Coster: The ionospheric impact of the October 2003 storm event on WAAS, *Proc. ION GNSS* (2004) pp. 1298–1307
- 39.76 M.V. Codrescu, C. Negrea, M. Fedrizzi, T.J. Fuller-Rowell, A. Dobin, N. Jakowski, H. Khalsa, T. Matsuo, N. Maruyama: A real-time run of the coupled thermosphere ionosphere plasmasphere electrodynamics (CTIPE) model, *Space Weather* **10**(2), 1–10 (2012)
- 39.77 J.M. Forbes: Planetary waves in the thermosphere-ionosphere system, *J. Geomagn. Geoelectr.* **48**, 91–98 (1996)
- 39.78 M.A. Abdu, T.K. Ramkumar, I.S. Batista, C.G.M. Brum, H. Takahashi, B.W. Reinisch, J.H.A. Sobral: Planetary wave signatures in the equatorial atmosphere-ionosphere system, and mesosphere-E- and F-region coupling, *J. Atmos. Solar-Terr. Phys.* **68**(3–5), 509–522 (2006)
- 39.79 P. Mukhtarov, B. Andonov, C. Borries, D. Pancheva, N. Jakowski: Forcing of the ionosphere from above and below during the Arctic winter of 2005/2006, *J. Atmos. Solar-Terr. Phys.* **72**(2/3), 193–205 (2010)
- 39.80 C. Borries, P. Hoffmann: Characteristics of F2-layer planetary wave-type oscillations in northern middle and high latitudes during 2002 to 2008, *J. Geophys. Res. Space Phys.* **115**(A11), 1–9 (2010)
- 39.81 M. Hernández-Pajares, J.M. Juan, J. Sanz: Medium-scale traveling ionospheric disturbances affecting GPS measurements: Spatial and temporal analysis, *J. Geophys. Res. Space Phys.* **111**(A7), 1–13 (2006)
- 39.82 K. Tsybulya, N. Jakowski: Medium- and small-scale ionospheric irregularities detected by GPS radio occultation method, *Geophys. Res. Lett.* **32**(A06302), 1–11 (2005)
- 39.83 E.L. Afraimovich: First GPS-TEC evidence for the wave structure excited by the solar terminator, *Earth Planets Space* **60**, 895–900 (2008)
- 39.84 M. Hernández-Pajares, J.M. Juan, J. Sanz, A. Aragón-Ángel: Propagation of medium scale traveling ionospheric disturbances at different latitudes and solar cycle conditions, *Radio Sci.* **47**(RS0K05), 1–22 (2012)
- 39.85 S. Lejeune, G. Wautelet, R. Warnant: Ionospheric effects on relative positioning within a dense GPS network, *GPS Solutions* **16**(1), 105–116 (2012)
- 39.86 E. Calais, J.B. Minster: GPS detection of ionospheric perturbations following the January 17, 1994, Northridge Earthquake, *Geophys. Res. Lett.* **22**(9), 1045–1048 (1995)
- 39.87 V. Ducic, J. Artru, Ph. Lognonné: Ionospheric remote sensing of the Denali Earthquake Rayleigh surface waves, *Geophys. Res. Lett.* **30**(18), 1–4 (2003)
- 39.88 N. Jakowski, V. Wilken, K. Tsybulya, S. Heise: Search of earthquake signatures from ground and space based GPS measurements. In: *Observation of the Earth System from Space*, ed. by J. Flury, R. Rummel, C. Reigber, M. Rothacher, G. Boedecker, U. Schreiber (Springer, Berlin 2006) pp. 43–53
- 39.89 J. Artru, V. Ducic, H. Kanamori, P. Lognonné, M. Murakami: Ionospheric detection of gravity waves induced by tsunamis, *Geophys. J. Int.* **160**(3), 840–848 (2005)
- 39.90 J.-Y. Liu, Y.-B. Tsai, K.-F. Ma, Y.-I. Chen, H.-F. Tsai, C.-H. Lin, M. Kamogawa, C.-P. Lee: Ionospheric GPS total electron content (TEC) disturbances triggered by the 26 December 2004 Indian Ocean tsunami, *J. Geophys. Res. Space Phys.* **111**(A05303), 1–4 (2006)
- 39.91 S. Pulinets, K. Boyarchuk: *Ionospheric Precursors of Earthquakes* (Springer, Berlin 2005)
- 39.92 J.Y. Liu, Y.J. Chuo, S.J. Shan, Y.B. Tsai, Y.I. Chen, S.A. Pulinets, S.B. Yu: Pre-earthquake ionospheric anomalies registered by continuous GPS TEC measurements, *Ann. Geophys.* **22**(5), 1585–1593 (2004)
- 39.93 T. Dautermann, E. Calais, J. Haase, J. Garrison: Investigation of ionospheric electron content variations before earthquakes in southern California, 2003–2004, *J. Geophys. Res. Solid Earth* **112**(B2), 1–20 (2007)

Reflectometry

40. Reflectometry

Antonio Rius, Estel Cardellach

This chapter discusses the use of properties of global navigation satellite system (GNSS) signals after their reflection on the Earth's surface. Global navigation satellite system reflectometry (or GNSS-R) is a multistatic radar that uses the GNSS constellations to extract information on the properties of the reflecting surfaces. Experiments have demonstrated that useful information can be extracted from such reflected signals. GNSS-R instruments have been installed in ground and coastal platforms, aircraft, stratospheric balloons, and spacecrafts. As a natural consequence it has been proposed by space agencies for the deployment of dedicated space missions. In the first part of this chapter the properties of the GNSS reflected signals on different components of the Earth's surface are discussed, and the technical principles sustaining different types of GNSS-R instruments are presented. The second part of this chapter presents methods to retrieve geophysical information from the GNSS-R signals, results obtained in different experiments and plans for future space missions.

40.1	Receivers	1164
40.1.1	GNSS-R Receivers	1165
40.2	Models	1167
40.2.1	Delay-Doppler Coordinates	1167
40.2.2	The Ambiguity Function	1167
40.2.3	The Noiseless Waveform Model	1169
40.2.4	Floor Noise Model	1169
40.2.5	Maximum Coherence Averaging Interval	1170
40.2.6	Speckle Noise	1171
40.2.7	Observed versus Modeled Waveforms	1171
40.3	Applications	1172
40.3.1	Sea Surface Altimetry	1173
40.3.2	Sea Surface Scatterometry	1175
40.3.3	Sea Surface Permittivity	1177
40.3.4	Cryosphere: Ice and Snow	1177
40.3.5	Land: Soil Moisture and Vegetation	1181
40.4	Spaceborne Missions	1182
	References	1183

GNSS reflectometry (GNSS-R) is an emerging technique aimed at inferring geophysical properties by means of analyzing the GNSS signals reflected off the Earth's surface. It thus exploits the remote sensing capabilities of the GNSS for Earth monitoring, in the form of a multiple bistatic radar (radar in which the transmitter and the receiver are at significantly distant locations), also called multistatic capabilities. The multistatic nature of the concept is based on the fact that each visible GNSS transmitter reflects onto a different area. If the GNSS-R receiver is at a sufficiently high altitude, this represents a set of simultaneous remote sensing observations covering a wide geographical zone, providing synoptic-view capabilities to the technique. Originally suggested in [40.1] as a system to perform mesoscale altimetry, it was later also identified as a scatterometric concept for ocean surface wind applications [40.2]. A review on GNSS-R can be found in [40.3, Chaps. 8–11].

Since then, a large number of experiments have been conducted. These experiments have used dedicated GNSS-R or software receivers, given that in most of the Earth's surfaces, and in particular over the ocean, the reflection process is essentially diffuse, inducing random and frequent shifts in the total phase of the electromagnetic field that impede standard GNSS tracking algorithms. Most of these GNSS standard applications obtain the primary observables, phases and pseudoranges, around the maximum of the cross-correlation between the GNSS signals and its mathematical replica, as explained in Chap. 14.

In contrast, the GNSS-R techniques extract the geophysical information from a wider extension of the cross-correlation function. These measurements are the primary observables in the GNSS-R case.

Reflectometry is a special form of multipath, but in most cases it is so extreme and diffuse that it cannot be modeled and analyzed as near-field multipath presented

in Chap. 15. Depending on the GNSS-R applications, the observables are linked to different aspects of the correlation function. While GNSS-R altimetry requires delay measurements, in the GNSS-R scatterometry case the observables are embedded in the shape of the distorted waveform.

40.1 Receivers

In this section we present the basic concepts used to collect the GNSS-R observables at the receiver level. Chaps. 13 and 14 of this Handbook present receiver architectures and signal processing techniques for GNSS navigation. In general, these do not work well for diffusely scattered signals that are characterized by low signal-to-noise ratio (SNR), signal fading events, and random phase behavior. This section therefore compiles basic information on architecture and signal processing adapted to GNSS-R.

We assume that a single-GNSS transmitted signal is received at a GNSS-R receiver through different paths, as sketched in Fig. 40.1:

- The direct path connects directly the transmitter T and the receiver R as in the standard GNSS applications, and
- The reflected path establishes the connection T-R after its reflection on a surface Σ .

This setup is similar to the one encountered in the *Lloyd's mirror* two-beam interference experiment, where a mirror creates a *virtual coherent image* of a source [40.4]. A difference is that in this chapter we assume that the surface Σ has *spatial incoherence*: (a) any pair of points of the source are statistically independent, and (b) a transmitted signal after its reflection on the incoherent surface Σ produces, at large distances, a coherent radiation that will interfere with the direct signal.

The primary GNSS-R observable is the *mutual coherence*, or *cross-correlation*, of these two separate GNSS signals created coherently by a single GNSS transmitter and collected by a single instrument, a GNSS-R receiver. Reference [40.4] covers in detail the study of the correlation functions of partially coherent light beams in the optical domain, and [40.5] apply these concepts to radio waves.

A GNSS-R receiver based on the measurement of the coherence of the reflected and direct beams will be termed interferometric or codeless. If a model of the code of the transmitted signal is known, the direct signal could be substituted by its known functional representation. We will use the terms clean-replica or code to

The formulation of the GNSS-R observables is given in Sect. 40.1; models for the reflected signals and their noise components are compiled in Sect. 40.2; a set of GNSS-R applications will be described in Sect. 40.3, and finally various GNSS-R spaceborne missions are presented in Sect. 40.4.

refer to this case. We should mention that the clean-replica GNSS-R receivers are conceptually close to the standard GNSS navigation receivers [40.6], while the interferometric receivers are close to the first geodetic quality codeless Global Positioning System (GPS) receivers [40.7] and the Very Long Baseline Interferometers [40.8].

The receiver can be understood as a system that collects signals and transforms them into other signals following specified procedures. We can assume that these signals can be modeled using mathematical functions expressing a statistical relation with experimental data.

Within this chapter, procedures and models will be indicated using the following conventions:

- When a signal B is the result of a process A , we will use the notation $B := A$.
- When we consider that a mathematical function A is an approximation of the dataset B , we will express such relation as $B \approx A$.

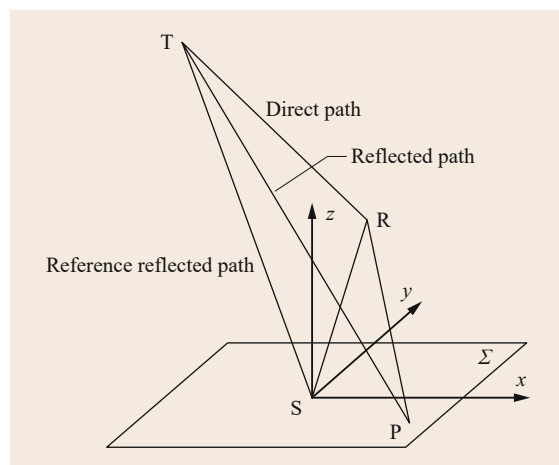


Fig. 40.1 Geometry of a GNSS reflection, and conventions used along this chapter. A signal transmitted at T is collected at R through direct and reflected paths. Point P is a generic point of the reflecting surface Σ . S is a point taken as a reference to compute the relative delays between the reflected and the direct signals as recorded in R

40.1.1 GNSS-R Receivers

We consider initially the interferometric GNSS-R case, as sketched in Fig. 40.1, where only one member T of the GNSS constellations transmits a band-limited radio frequency (RF) signal. Two antennas placed at the receiver position R collect samples of this signal. One, the uplooking antenna, provides samples of the signal traveling through the direct path. The other, the down-looking antenna, provides samples of the same signal after its reflection on the points P of surface Σ , traveling through a diversity of reflected paths.

The collected signals are real-valued functions. In this chapter we will use its associated analytic function, as used in communication theory. This allows a simpler mathematical formulation (see, for instance, [40.9] or [40.5, App. 3.1]).

The direct samples $V_D(t)$, where t is the time as measured by the instrument clock, will be represented by the functional model

$$V_D(t) \approx A_D(t)e^{+2\pi j\nu_0 t}, \quad (40.1)$$

where $A_D(t)$ is a complex covariance-stationary stochastic process that modulates a tone at the frequency ν_0 .

We assume that the transmitted signal arrives also to the receiver after its reflection on the surface Σ as function of the time t , as the complex signal $V_R(t)$,

$$V_R(t) \approx A_R(t - \tilde{\tau}(t))e^{+2\pi j\nu_0(t - \tilde{\tau}(t))}, \quad (40.2)$$

where $A_R(t)$ is a complex covariance-stationary stochastic process and $\tilde{\tau}(t)$ is a model that predicts the relative delay between a signal reflected in a reference point S and the direct signal. In the following discussion the reference point selected is the specular point. Note that the processes $A_R(t)$ and $A_D(t)$ at the input of the GNSS-R receivers are filtered versions of the nominal values.

To align the direct signal, we delay the direct signal using the model $\tilde{\tau}(t)$, to obtain the delay compensated signal defined as

$$\begin{aligned} V_D^c(t) &:= V(t - \tilde{\tau}(t)), \\ V_D^c(t) &\approx A_D(t - \tilde{\tau}(t))e^{+2\pi j\nu_0(t - \tilde{\tau}(t))}. \end{aligned} \quad (40.3)$$

Because the granularity of the sampling process is too coarse, the alignment requires some form of interpolation. This may be accomplished in several ways. Time delay interpolation procedures can be found in [40.10] and [40.11] and their frequency-domain counterpart in [40.8].

Once both signals are aligned, we compute the coherence, or cross-correlation, of both signals using

$$\begin{aligned} \Gamma_{RD}(t_c, \tau) &:= \langle V_R(t + \tau)V_D^c(t) \rangle_{T_c} \\ &:= \frac{1}{T_c} \int_{T_c} V_R(t + \tau)V_D^c(t) dt \\ &:= V_R(t) \star V_D^c(t), \end{aligned} \quad (40.4)$$

where the asterisk $*$ indicates the complex conjugate value, t_c is the epoch associated to the coherent averaging interval T_c , τ is the lag delay variable, $\langle \dots \rangle_T$ denotes the statistical average during the interval T , and \star is the correlation symbol.

Using (40.1) and (40.3), and assuming that the cross-correlation $\Gamma_{RD}(t_c, \tau)$ is nearly time invariant, we obtain the corresponding mathematical model as

$$\Gamma_{RD}(t_c, \tau) \approx \langle A_R(t + \tau)A_D^*(t) \rangle_{T_c}. \quad (40.5)$$

The intrinsic variations of the surface Σ and the movements of the transmitter and the receiver will limit the size of the interval T_c , according to the van Cittert–Zernike theorem (e.g., [40.4] and [40.5]). An expression to compute the order of magnitude of T_c is given in Sect. 40.2.5.

To reduce thermal and speckle noise (see Sect. 40.2.6), the receiver delivers the waveforms W_{RD} computed as the average of $|\Gamma_{RD}(t_c, \tau)|^2$ during an incoherent averaging interval T_a

$$\begin{aligned} W_{RD}(t_a, \tau) &:= \langle |\Gamma_{RD}(t_c, \tau)|^2 \rangle_{T_a}, \\ W_{RD}(t_a, \tau) &\approx \langle |A_R(t + \tau)A_D^*(t)|^2 \rangle_{T_c}. \end{aligned} \quad (40.6)$$

Here t_a is the time associated with the averaging interval T_a . Note that these products will depend on the choice of $\tilde{\tau}$ as well as the duration of the intervals T_c and T_a .

This formulation can be applied directly to the clean-replica case, assuming that the power of the signal arriving through the direct link is noiseless and has unit power. It should also be noted here that the amplitude of the direct $A_D(t)$ and the reflected $A_R(t)$ signals at the correlator input depend linearly on $\sqrt{P_T}$, where P_T is the transmitted power and consequently the GNSS-R waveforms $W_{RD}(\tau)$ will depend quadratically on P_T . In the clean-replica approach V_D is a replica or template with unity power at the correlator input. In this case the thermal input noise in the direct link is removed, and the output of the receiver will depend only linearly on the transmitted power P_T . The normalized interferometric GNSS-R waveforms defined as

$$W_{RD}^n(t_a, \tau) := \frac{W_{RD}(t_a, \tau)}{P_T} \quad (40.7)$$

will be the equivalent to the clean-replica waveforms.

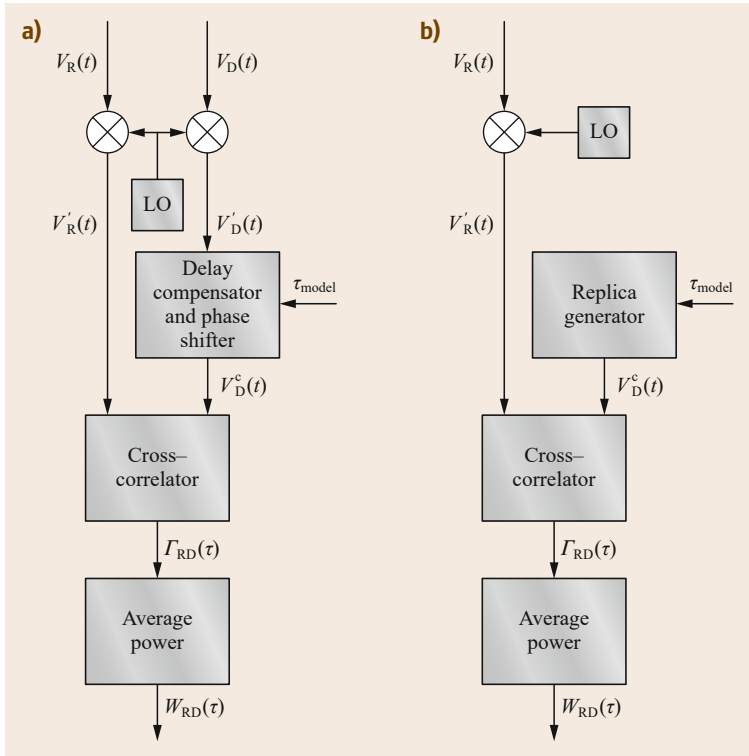


Fig. 40.2a,b Sketch of interferometric (a) and clean-replica (b) GNSS-R Receivers

GNSS-R Receivers with Downconversion

The waveforms computed as per (40.6) require the direct sampling of the signal at the radio frequency (RF) band. As in the standard GNSS navigation receivers, it is preferred to downconvert the signal to baseband. We provide now a possible procedure.

An interferometric GNSS-R receiver with the correlation performed with lower sampling rate is repre-

sented conceptually at the left of Fig. 40.2. A clean-replica GNSS-R receiver is shown at the right in the same figure. The interferometric GNSS-R receiver accepts the direct $V_D(t)$ and reflected $V_R(t)$ signals and transforms them through the following steps:

Downconversion: The spectrum of both signals is shifted coherently from the RF band to a lower band. Assuming that the downconversion frequency is ν_0 , we have

$$\begin{aligned} V'_D(t) &:= V_D(t)e^{-2\pi j\nu_0 t}, \\ V'_D(t) &\approx A_D(t), \end{aligned} \quad (40.8)$$

$$\begin{aligned} V'_R(t) &:= V_R(t)e^{-2\pi j\nu_0 t}, \\ V'_R(t) &\approx A_R(t - \tilde{\tau}(t))e^{-2\pi j\nu_0 \tilde{\tau}(t)}. \end{aligned} \quad (40.9)$$

Delay and Phase Compensation: The remaining transformations are performed only over the direct signal. First, the direct downconverted signal is delayed according to the model $\tilde{\tau}(t)$ to obtain the shifted direct

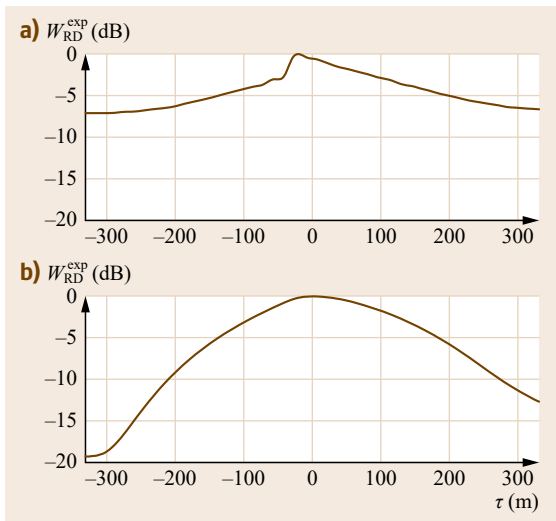


Fig. 40.3a,b Examples of waveforms W_{RD} obtained simultaneously during a flight experiment. (a) Corresponds to an interferometric waveform, of a composite signal that included C/A, P(Y) and M codes. (b) Panel presents a C/A clean-replica waveform. Both waveforms have been divided by their maximum value ◀

downconverted signal V_D'

$$\begin{aligned} V_D''(t) &:= V_D'(t - \tilde{\tau}(t)), \\ V_D''(t) &\approx A_D(t - \tilde{\tau}(t)). \end{aligned} \quad (40.10)$$

Subsequently, this shifted direct downconverted signal V_D'' is rotated by a phase angle $-2\pi\nu_0\tilde{\tau}(t)$ to obtain the downconverted and aligned signal $V_D^c(t)$

$$\begin{aligned} V_D^c(t) &:= V_D''(t)e^{-2\pi j\nu_0\tilde{\tau}(t)}, \\ V_D^c(t) &\approx A_D(t - \tilde{\tau}(t))e^{-2\pi j\nu_0\tilde{\tau}(t)}. \end{aligned} \quad (40.11)$$

Comparison of the signal models given in (40.9) and (40.11) shows that the signal $V_D^c(t)$ has been aligned

with the compensated reflected signal $V_R'(t)$. Consequently the waveforms obtained with a receiver with a baseband correlator will be equivalent to those obtained with a direct sampling correlation receiver.

In Fig. 40.3 we present two waveforms obtained simultaneously with interferometric and clean-replica receivers placed on an aircraft flying at 3000 m altitude. The accumulation periods were $T_c = 1$ ms and $T_a = 20$ s. In this example it can be noted that the dynamic ranges of these waveforms are 7 dB and 19 dB respectively. Also it should be noted that the leading edge of the interferometric waveform presents a steeper slope than the clean-replica waveform. Reference [40.12] provides a detailed account of this experiment.

40.2 Models

The purpose of this section is to outline the procedures to model the observed waveforms in terms of experimental parameters: positions, velocities, and roughness of the reflecting surface. As each point P in the surface Σ will contribute signals with different delay and Doppler shifts to the waveform, we start assigning values for the expected delay and Doppler to each point. The second step will be to assemble the waveform model.

40.2.1 Delay-Doppler Coordinates

To build the model $\tilde{\tau}$ we assume, as a first approximation, that the positions of the transmitter T, the receiver R and the reflecting surface Σ are known functions of time. At the epoch t , as measured by the receiver clock, we define the relative delay model $\tilde{\tau}(P)$ (expressed in distance units) as

$$\tilde{\tau}(P) := \rho(T, P) + \rho(P, R) - \rho(T, R), \quad (40.12)$$

where $\rho(A, B)$ is the geometric range between two points A and B at this epoch.

Note that a more accurate model should take into consideration that at the epoch t the recorded signals $V_D(t)$ and $V_R(t)$ were transmitted when the transmitter was at $T(t - \delta\tau_D)$ and $T(t - \delta\tau_R)$, where $\delta\tau_D$ and $\delta\tau_R$ are the propagation time from the transmitter to the receiver through the direct and the reflected beam respectively. For the purposes of aligning the direct and the reflected signals within the correlation window the model (40.12) is sufficient.

We will take as the reference the specular point S where $\tilde{\tau}(P)$ is a minimum. The quantity

$$\Delta\tau(P) := \tilde{\tau}(P) - \tilde{\tau}(S) \quad (40.13)$$

will be referred as the specular relative delay. The points P of the surface Σ with $\Delta\tau(P) = \Delta\tau_0$ and $\Delta\tau_0 > 0$ define the isodelay curve $\Delta\tau_0$.

The effects of the changing geometry can be analyzed taking into account the velocities of the points T, P and R. In this case, the derivative of the relative delay model, as stated in (40.12), is

$$\frac{d\tilde{\tau}(P)}{dt} = \frac{d\rho(T, P)}{dt} + \frac{d\rho(P, R)}{dt} - \frac{d\rho(T, R)}{dt}. \quad (40.14)$$

This variation expresses the difference between the range rate of the signal reflected at P and the direct signal, or Doppler shift

$$v(P) = -\frac{v_0}{c} \frac{d\tilde{\tau}(P)}{dt}. \quad (40.15)$$

Here c is the speed of light in vacuum and the quantity

$$\Delta v(P) = v(P) - v(S) \quad (40.16)$$

is the relative Doppler shift experienced by two signals reflected at P and S.

The points P of the surface Σ where $\Delta v(P) = \Delta v_0$ define the iso-Doppler lines Δv_0 . As a reference, in Fig. 40.4b we have represented a set of isodelay and iso-Doppler lines for a receiver placed at 3 km altitude with a horizontal velocity of 55 m/s.

40.2.2 The Ambiguity Function

To model the cross-correlation defined in (40.4) and (40.6), we start with the contribution of the reflected signal in the point P, whose delay-Doppler coordinates

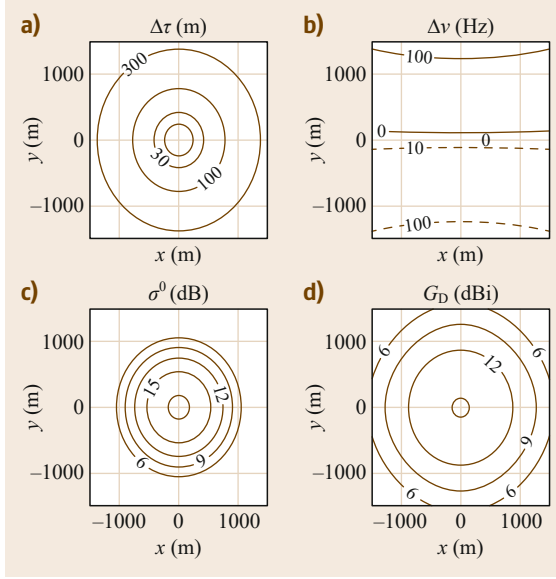


Fig. 40.4a–d Sketches of isolines in the reflecting surface corresponding to one experiment where a receiver was placed at 3 km altitude flying at 55 m/sec. The functions represented are relative delay $\Delta\tau(\mathbf{P})$ (a), relative frequency $\Delta\nu(\mathbf{P})$ (b), bistatic scattering coefficient $\sigma_{pq}^0(\mathbf{P})$ (c) and antenna gain $G_D(\mathbf{P})$ (d)

are $\Delta\tau(\mathbf{P})$ and $\Delta\nu(\mathbf{P})$

$$A_R(\mathbf{P}, t + \tau) \approx \sqrt{P_R(\mathbf{P})} \hat{A}_D(t + \delta\tau) \times e^{-2\pi j \delta\nu t}, \quad (40.17)$$

where τ is the correlator delay, $\delta\tau = \tau - \Delta\tau(\mathbf{P})$, $\delta\nu = \Delta\nu(\mathbf{P})$, $P_R(\mathbf{P})$ is the average power of the collected reflected signal coming from the point \mathbf{P} , and $\hat{A}_D(t) = A_D(t)/P_D$ where P_D is the average power of the direct signal.

The contribution of each point \mathbf{P} of the reflecting surface to the correlator output will be

$$\begin{aligned} \Gamma_{RD}(\mathbf{P}, \tau) &\approx \sqrt{P_R(\mathbf{P})} \sqrt{P_D} \\ &\times \left\langle \hat{A}_D(t + \delta\tau) \hat{A}_D^*(t) e^{-2\pi j \delta\nu t} \right\rangle_{T_c} \\ &\approx \sqrt{P_R(\mathbf{P})} \sqrt{P_D} \chi_A(\delta\tau, \delta\nu), \end{aligned} \quad (40.18)$$

where χ_A is the function

$$\chi_A(\tau, \nu) := \frac{1}{T_c} \int_{T_c} \hat{A}_D(t + \delta\tau) \hat{A}_D^*(t) e^{-2\pi j \delta\nu t} dt. \quad (40.19)$$

The contribution of the signal reflected at point \mathbf{P} to the waveform will be

$$W_{P, RD}(\tau) := P_D P_R(\mathbf{P}) |\chi_A(\delta\tau, \delta\nu)|^2. \quad (40.20)$$

The function $|\chi_A(\delta\tau, \delta\nu)|^2$ in the above equation is usually called the narrowband radar ambiguity function associated to $\hat{A}_D(t)$ [40.13]. It represents the power radar return of a moving target whose delay and Doppler shift with respect to the nominal values are $\delta\tau$ and $\delta\nu$. The term *ambiguity* refers to the uncertainty of the delay and Doppler radar measurements. The function $|\chi_A(\delta\tau, \delta\nu)|^2$ can be estimated from known correlation properties of the transmitted signals, or determined experimentally.

Two properties of the ambiguity functions are relevant to determine its overall structure. Along the axis $\delta\nu = 0$ the ambiguity function is the squared autocorrelation of the process $\hat{A}_D(t)$

$$|\chi_A(\delta\tau, 0)|^2 := \left| \frac{1}{T_c} \int_{T_c} \hat{A}_D(t + \delta\tau) \hat{A}_D^*(t) dt \right|^2. \quad (40.21)$$

Within the GNSS-R literature, this value is commonly denoted as

$$|A(\delta)|^2 \equiv |\chi_A(\delta\tau, 0)|^2. \quad (40.22)$$

In the case that $\hat{A}(t) \hat{A}^*(t) = 1$, the ambiguity function along the axis $\delta\tau = 0$ is

$$|\chi_A(0, \delta\nu)|^2 := |\text{sinc}(\pi \delta\nu T_c)|^2, \quad (40.23)$$

where the sinc function is defined as

$$\text{sinc}(x) = \frac{\sin(x)}{x}. \quad (40.24)$$

In Fig. 40.5, we represent two examples of $|\chi(\delta\tau, 0)|^2$; a GPS C/A signal and a composite C/A-, P(Y)- and M-code modulated signal. The assumed effective isotropic radiated power (EIRP), defined as the product of the transmitted power P_T and the gain of the transmitter antenna G_T^D in the direction of the receiver, for each component as the its modulations are given in Table 40.1. We have assumed that the precorrelation bandwidth is 24 MHz.

Table 40.1 Signals used in the simulation. See Chap. 4 for a description of the individual modulation types

Signal	EIRP	Modulation	Reference
C/A	24.0 dBW	BPSK(1)	[40.6]
P(Y)	21.3 dBW	BPSK(10)	[40.6]
M	25.3 dBW	BOC(10,5)	[40.14]

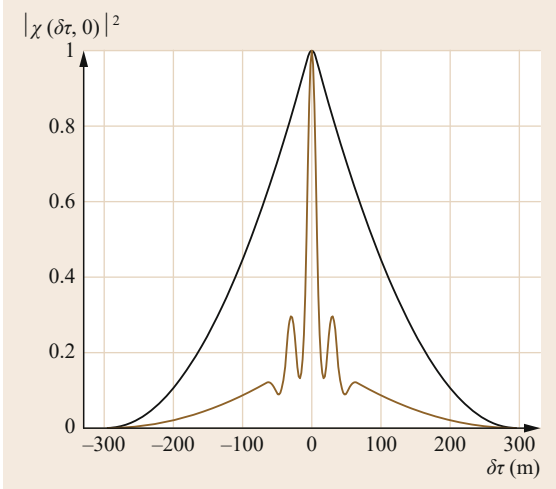


Fig. 40.5 Examples of $|\chi(\delta\tau, 0)|^2$: *black line* corresponds to a GPS C/A signal, and the *brown line* to a GPS signal composite of C/A-, P(Y)- and M-codes

40.2.3 The Noiseless Waveform Model

Adding all the contributions $W_{RD}(P, \tau)$, defined in (40.20) from all the points P of the surface Σ we obtain a model for the waveform

$$W_{RD}(\tau) := P_D \int_{\Sigma} P_R(P) |\chi_A(\delta\tau, \delta\nu)|^2 d\sigma \quad (40.25)$$

and, according to the definition (40.7), the normalized waveform will read

$$W_{RD}^n(\tau) := \int_{\Sigma} P_R(P) |\chi(\delta\tau, \delta\nu)|^2 d\sigma. \quad (40.26)$$

The received power P_D in the (40.25) at the input of the correlator is related to the transmitter power by

$$P_D := P_T G_T^D G_R^D \left(\frac{\lambda}{4\pi R_{TR}} \right)^2, \quad (40.27)$$

where G_T^D and G_R^D are the transmitter and receiver antenna gains for the direct link, R_{TR} is the distance between transmitter T and receiver R, and λ is the signal wavelength. Similarly, the received power of the reflected signal is given by

$$P_R(P) := \frac{P_T G_T(P) G_R(P) \sigma_{pq}^0(P) \lambda^2}{(4\pi)^3 R_T(P)^2 R_R(P)^2}. \quad (40.28)$$

Here, $\sigma_{pq}^0(P)$ is the bistatic scattering differential coefficient. For each point P of the surface Σ , this coefficient is defined as the ratio of the power scattered

towards the receiver R with polarization state q to the power incident with polarization state p , per unit area (see [40.15, p. 463] for details). This ratio depends on the probability density function (PDF) of the surface roughness slopes. Reference [40.16, App. B], presents a widely accepted $\sigma_{pq}^0(P)$ model, where the sea slopes PDF is a bivariate normal distribution, later introduced in Sect. 40.3.2.

As a reference, we have represented a set of isogain and iso- σ lines that model the values corresponding to the aforementioned experiment at the bottom of Fig. 40.4.

40.2.4 Floor Noise Model

In this section we present a simple model to describe the noise to be added to the noiseless waveforms defined in (40.25).

We assume now that the compensated recorded signals $V_D^c(t)$ and $V_R^c(t)$ are

$$\begin{aligned} V_D^c(t) &\approx v_D^c(t) + n_D(t), \\ V_R^c(t) &\approx v_R^c(t) + n_R(t), \end{aligned} \quad (40.29)$$

where $v_D^c(t)$ and $v_R^c(t)$ are the compensated noiseless signal terms and $n_D(t)$ and $n_R(t)$ the noise terms.

The coherence $\Gamma_{RD}(t_c, \tau)$, defined in (40.4), will read

$$\begin{aligned} \Gamma_{RD}(\tau) &:= (v_R^c + n_R) \star (v_D^c + n_D) \\ &= v_R^c \star v_D^c \\ &\quad + v_R^c \star n_D + n_R \star v_D^c + n_R \star n_D \end{aligned} \quad (40.30)$$

and the corresponding waveforms is given by (40.6)

$$W_{RD}(\tau) := W_{RD}^{\text{signal}}(\tau) + N, \quad (40.31)$$

where $W_{RD}^{\text{signal}}(\tau)$ is the noiseless waveform, modeled as per (40.25), (40.27), and (40.28), and the term N includes the contributions of the noise terms.

According to [40.17], the waveform noise term is given by

$$N = P_D \frac{k_B T_R}{T_c} + P_R \frac{k_B T_D}{T_c} + k_B T_R B \frac{k_B T_D}{T_c}, \quad (40.32)$$

where k_B is the Boltzmann constant, B is the pre-detection bandwidth, and T_c is the coherent averaging interval, already introduced. T_D and T_R are the effective noise temperatures of the direct and reflected signals. They are defined as

$$T = T^{\text{ant}} + (F - 1) 290 \text{ K}, \quad (40.33)$$

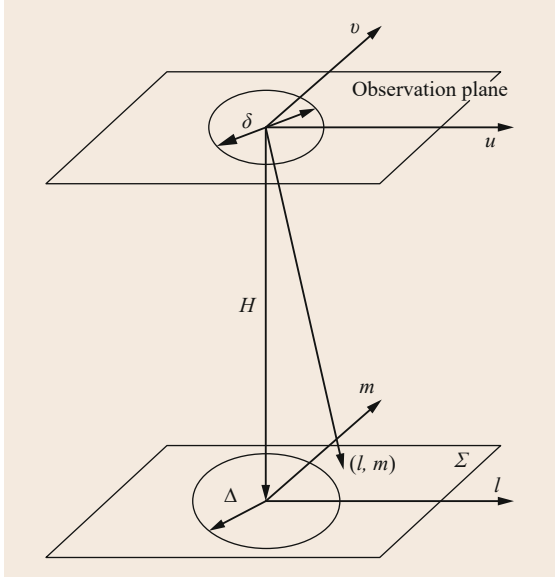


Fig. 40.6 Sketch to show the (u, v) and (l, m) coordinates used in the formulation of the van Cittert–Zernike theorem. The distribution of the radiation intensity in the Σ space and the coherence in the observation plane form a Fourier transform pair

where T^{ant} is the corresponding antenna temperature and F is the noise factor expressed in linear units.

In a clean-replica instrument $T_D = 0$, and the corresponding noise term will be

$$N^{\text{cr}} := P_D \frac{k_B T_R}{T_c}. \quad (40.34)$$

The quantities N and N^{cr} represent the effective waveform floor noise for the interferometric and clean-replica cases. Parts of waveforms with power below these quantities will not be observable.

To express the relation between the interferometric and clean-replica waveforms we introduce the interferometric noise factor η defined as

$$\begin{aligned} \eta &:= \frac{N}{N^{\text{cr}}} \\ &:= 1 + \frac{1 + (S/N)_R}{(S/N)_D}, \end{aligned} \quad (40.35)$$

where the power signal-to-noise ratios at the correlator input are defined as

$$\begin{aligned} (S/N)_D &:= \frac{P_D}{k_B T_D^n B}, \\ (S/N)_R &:= \frac{P_R}{k_B T_R^n B}. \end{aligned} \quad (40.36)$$

As pointed out in [40.17], $\eta \rightarrow 1$ for $(S/N)_D \gg 1 + (S/N)_R$. In this case the noise in the interferometric waveforms do not increase substantially with respect to the clean-replica case. However, the situation is different in the low SNR regime. When $(S/N)_D \ll 1$, we have approximately

$$\eta \approx \frac{1}{(S/N)_D}, \quad (40.37)$$

i. e., the interferometric noise factor is inversely proportional to the signal-to-noise ratio of the direct signal.

40.2.5 Maximum Coherence Averaging Interval

In this section we attempt to provide an order of magnitude of the maximum coherence time T_c to be used as the coherent averaging interval.

We assume that an incoherent extended source is placed in the surface Σ (Fig. 40.6). The van Cittert–Zernike theorem ([40.4, 5] for background material, and [40.18] for its application to GNSS-R) states that the Fourier transform of the intensity of the source I is the coherence function Γ in the observation plane, assumed parallel to Σ

$$\Gamma(u, v) = \int_{\Sigma} I(l, m) e^{-j2\pi(ul+vm)} dl dm. \quad (40.38)$$

Here, u and v are spatial coordinates in the observation plane, expressed in wavelength λ units, and (l, m) are the direction cosines defined respect to the (u, v) axis.

Assume now that the distance between both planes is H . A section of the surface Σ of linear dimensions Δ , when expressed in terms of the direction cosine has the size Δ/H . The Fourier transform relationship implies that the dimensions of its image in the observation plane expressed in linear units will be $\delta = \lambda H / \Delta$.

The radius of the section of the surface Σ limited by the isodelay line τ , expressed in linear units, is $\sqrt{2\tau H}$. Then the dimensions in the observation plane produced by such a section will be

$$\delta = \frac{\lambda \sqrt{H}}{2\sqrt{2\tau}}. \quad (40.39)$$

If the receiver is moving at velocity v_R , the samples obtained within an interval

$$T_c < \frac{\delta}{v_R} = \frac{\lambda \sqrt{H}}{2\sqrt{2\tau} v_R} \quad (40.40)$$

will show coherence.

In the case of a coastal experiment, $v_R = 0$, and the previous expression should be substituted by a constant value expressing the intrinsic variability of the sea surface. Reference [40.19] indicates that this intrinsic variability is on the order of 100 msec.

40.2.6 Speckle Noise

Equation (40.20) expresses the contribution to the waveform $W_{RD}(\tau)$ of the different elements of surface $d\sigma$ of the reflecting surface Σ . The quantity $W_{RD}(\tau)$ for a particular value of τ will be built by the contributions of a large number N of scatterers placed in the proximity of the isodelay line labeled τ . Each scatterer n will contribute with a signal $V_n e^{j\phi_n}$ to the sum

$$V(\tau) = V_e e^{j\phi} = \sum_{n=1}^{N_s} V_n e^{j\phi_n}. \quad (40.41)$$

The amplitude of this sum, V_e , has a Rayleigh distribution, and its phase ϕ is uniformly distributed in the range $(0, 2\pi)$. The probability distribution associated to the power $P(\tau) = V^2(\tau)$ is exponential

$$p(P) = \begin{cases} \bar{P} e^{-P/\bar{P}} & \text{if } P \geq 0 \\ 0 & \text{if } P < 0, \end{cases}$$

where \bar{P} is the mean value and the standard deviation of the distribution. This large dispersion is reduced by

Table 40.2 Main instrument and experiment parameters. Labels D and R corresponds to the direct and reflected signal respectively. **RHCP**: right-hand circular polarized, **LHCP**: left-hand circular polarized

Variable	Value	Symbol
Antenna gain (D) (R)	15 dBi	G_R, G_D
Polarization (D)	RHCP	p
Polarization (R)	LHCP	q
Antenna temp. (D)	10 K	T_D^{ant}
Antenna temp. (R)	200 K	T_R^{ant}
Noise factor (D) (R)	3 dB	F
RF bandwidth (D) (R)	24 MHz	B
Coherent ave. time	1 msec	T_c
Incoherent ave. time	20 sec	T_a
Mean-squared slope	≈ 0.010	MSS
Incidence angle	$\approx 0^\circ$	θ
Height receiver	≈ 3 km	h_{RS}
Velocity receiver	≈ 55 m/s	v_R

averaging many individual waveforms gathered during the incoherent averaging interval T_a .

40.2.7 Observed versus Modeled Waveforms

In this section we analyze the waveforms shown in Fig. 40.3 in terms of the signal and noise models described in the two previous sections. They were acquired simultaneously during an aircraft experiment

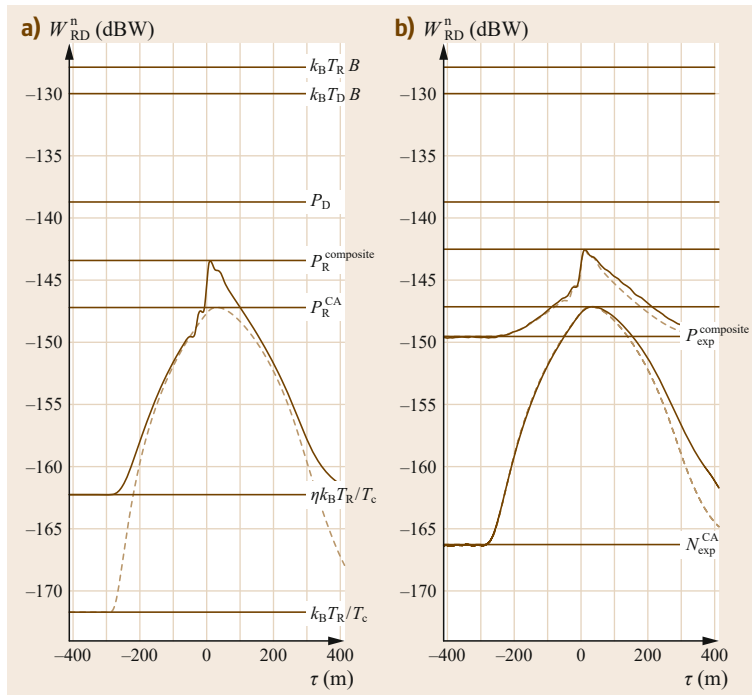


Fig. 40.7 (a) Modeled waveforms for the clean-replica (dashed line) and interferometric (brown) cases. (b) Observed corresponding waveforms (brown), where its total power has been adjusted using the modeled values, and with the same dynamic range as from the experimental results

Table 40.3 Modeled power levels and interferometric noise factor for the aircraft experiment

Signal/Variable	dBW	Symbol
Pre-correlation noise (D)	-130.2	$k_B T_D B$
Pre-correlation noise (R)	-127.9	$k_B T_R B$
Peak composite waveform (D)	-138.7	$P_D^{\text{composite}}$
Peak composite waveform (R)	-143.4	$P_R^{\text{composite}}$
Peak CA waveform (R)	-147.2	P_R^{CA}
Floor noise composite	-160.9	N
Floor noise CA	-171.7	N^{cr}
Interferometric noise factor	10.8	η

described and analyzed in [40.12]. The relevant parameters describing the experiment are indicated in Table 40.2. We have included in this table the approximate values for the mean squared slopes (MSSs), incidence angle of the signal θ , and the height h_{RS} and the nearly horizontal velocity v_R of the receiver needed for modeling the corresponding waveforms.

The model $\hat{\tau}(t)$, needed to align the direct and the reflected signals, has been obtained using GPS and Inertial Measuring Unit observables acquired during the experiment, and International GNSS Service (IGS) GPS precise orbits.

In Fig. 40.7 we have represented, in the left panel, interferometric and a clean-replica waveforms modeled

Table 40.4 Experimental floor power levels and interferometric noise factor

Signal/Variable	dBW	Symbol
Floor noise composite	-149.5	N_{exp}
Floor noise CA	-166.2	$N_{\text{exp}}^{\text{cr}}$
Interferometric noise factor	16.7	η_{exp}

as per (40.31), and normalized as before. In Table 40.3 we have indicated different values of the power.

In the left panel of Fig. 40.7, we have represented model and experimental waveforms. We have adjusted the position of observed waveforms to the corresponding position of the model. The floor noise of this real waveforms has been used to define the floor noise for the composite and C/A modeled waveforms. These experimental floor noise values $N_{\text{exp}}^{\text{composite}}$ and $N_{\text{exp}}^{\text{CA}}$ are given in Table 40.4.

The difference of 6 dB between the experimental and modeled interferometric noise factor can be attributed to known model limitations in the parametrization used. Among others, these include:

- Losses before the low noise amplifiers
- Quantization noise
- Atmospheric attenuation, and
- Power transmitted in each code.

40.3 Applications

As it has been shown in the previous sections, GNSS reflectometry can be seen as a particular form of bistatic radar with capabilities for multistatic coverage.

The rich amount of available GNSS signals covering the entire globe (e.g., Fig. 40.8), and being transmitted at least at two frequencies of the L-band spectrum (all weather propagation), are some of the advantages of the technique. The density of observations that this concept can offer from space-based platforms potentially results in a high spatiotemporal resolution compared to some dedicated techniques. On the other hand, the fact that these signals have not been designed for remote sensing purposes makes them suboptimal in some aspects, especially when compared to dedicated remote sensing concepts. Some of the limiting factors are the transmitted powers, and the bandwidth and pulse resolution.

Despite these limitations, and because of the high and quick coverage of the Earth, the technique has potential to fill some gaps in the current Earth observational system. Examples of these gaps include meso-scale altimetry and scatterometry, both within a few days' time resolution and unaffected by the weather conditions (rain in particular).

The growing interest for these reflectometry techniques has boosted the development of dedicated instrumentation and experimental campaigns. Since the late 1990s and until October 2013, the amount of GNSS-R experiments conducted aboard aircrafts was greater than 270, executed with more than 20 different GNSS-R receivers. These experiments have been conducted in different geographic localizations, and they captured reflections off the ocean, bare soil, growing crops, lakes and rivers, sea ice, glaciers, mid-latitudes and Antarctic snow. Some of these datasets are available for research purposes at a web server [40.20] and documented in [40.21]. Moreover, the feasibility of receiving GNSS signals reflected off the Earth's surface from spaceborne platforms has also been proven. The first GNSS reflection captured from outer space was collected with the SIR-C instrument aboard the Space Shuttle [40.22]. A dedicated GNSS-R instrument with a moderate antenna gain (11.8 dBi) also gathered reflected signals off ocean, ice, and land surfaces from 700 km aboard one of the UK Disaster Monitoring Constellation (UK-DMC) satellites [40.23–25]. Sets of GNSS-R data obtained with the UK-DMC Low Earth

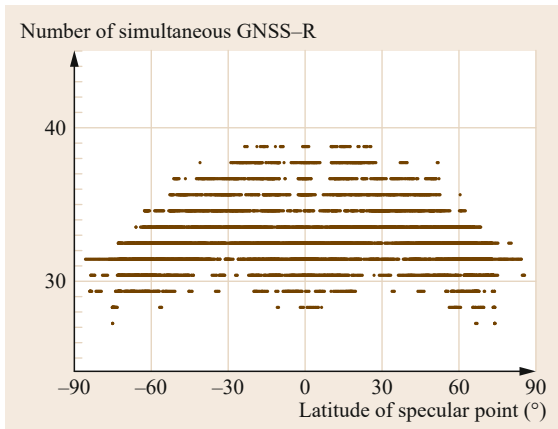


Fig. 40.8 Number of simultaneously reflected GNSS satellites as a function of the latitude coordinate of their specular point on the Earth's surface, and accumulated in one day of observations. Two GNSS constellations (GPS and GLONASS) have been assumed, status as in 18 March 2012. These values might be doubled once Galileo and BeiDou-3 are fully operational

Orbiters are also available at [40.26]. Some of the applications that have resulted out of these studies are summarized in the following sections.

40.3.1 Sea Surface Altimetry

The oceanographic applications were the first to be identified as potential targets of the GNSS-R technique. In fact, the concept was conceived for sea surface altimetry in 1993 [40.1], and shortly after as a system to measure sea surface roughness (scatterometry) [40.2].

Ocean altimetry is one of the most challenging applications of the GNSS-R. In order to achieve the levels of precision needed for scientifically valuable products, the system requires demanding signal-to-noise ratios (SNR), maximizing the use of the bandwidth, and preserving the multistatic nature of the concept. Altogether, it introduces complexity at the hardware level (antennas, beam-formers, signal processors). Despite this complexity, accurate enough GNSS-R altimetric measurements have the potential to fill gaps in the current observational system, especially for its dense and quick global coverage, enabling the capture of ocean signals such as tsunamis (e.g., [40.27]), eddies and other mesoscale oceanic features. A large fraction of the ocean's kinetic energy is associated with spatial scales that cannot be resolved with a single radar altimeter mission. Mesoscale variability is key to understanding large-scale circulation and climate variability. The assimilation of GNSS-R data into ocean circulation models was first studied in [40.28], showing positive

impact. Another benefit of the multistatic nature of the GNSS-R altimetry is its potential to resolve the sea surface slope (gradients in topography) in two dimensions, something that standard radar altimetric missions achieve only in the intersections of their ground tracks.

The scattering of GNSS signals off the sea surface tends to be diffuse, with little or no coherent part. The main consequence of diffuse scattering is that the phase of the reflected signal can barely be tracked, and the phase-delay observables cannot generally be obtained. As a consequence, the range measurements rely on group-delay measurements of lower precision. Therefore, this section focuses on group-delay ocean altimetry, while phase-delay altimetry is mostly used for ice applications (Sect. 40.3.4).

Chapter 14 shows that despite its higher accuracy, group-delay pseudoranges are far less precise than ranges measured with phase-delay observables, and describes how to extract the group-delay observables from the triangular GNSS waveform: the procedure can be summarized as the estimation of the delay of the peak of the triangle function resulting from cross-correlating the signals against their receiver synthesized replicas. This procedure is generally not suitable for reflected signals, since the waveform is strongly distorted by the severe and random multipath-like effect of reflections across a wide area on the surface; the glistening zone. The rougher the surface the wider the glistening zone, and thus the delay extension of multipath due to off-specular areas. This effect is illustrated in Fig. 40.9. The peak-delay estimator to extract the range of the reflected signals was nevertheless used in experiments over calm waters and low altitudes, such as in [40.29] or [40.30]. As described in Sect. 40.2.6, speckle is a major component of the noise, resulting in multiplicative dispersion following an exponential distribution of probability. Under these circumstances new approaches to identify the specular range within the distorted waveform are needed.

The procedure to identify the specular delay under diffuse scattering is composed by at least two steps: (1) sufficiently long incoherent averaging to reduce speckle; and (2) estimation of the specular delay. If the waveforms are obtained in real time (e.g., using a hardware receiver) some realignment of the waveforms along the delay axis might be required before integration, to correct for errors in the delay model applied by the receiver based on the real-time information. Integration without these corrections would result in blurred integrated waveforms. Two approaches are mainly used to estimate the delay location of the specular ray path. The first one found in the literature was based on fitting a model of the waveform, such as the one in (40.25). The model requires the assumption of

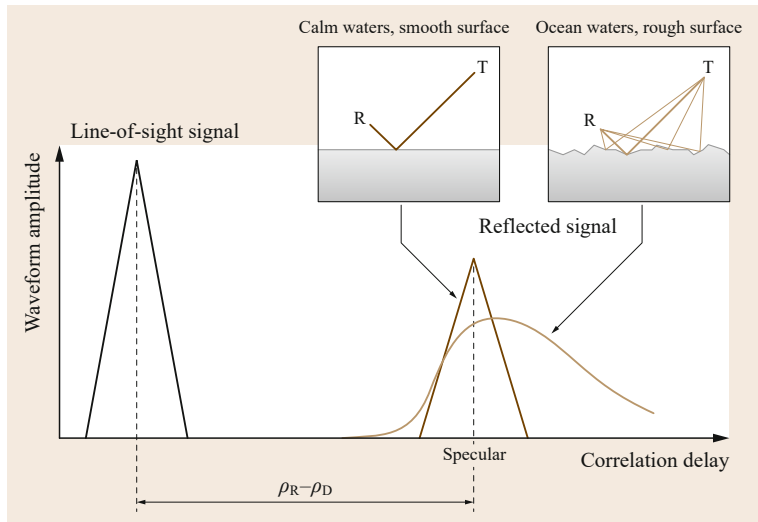


Fig. 40.9 As illustrated in Fig. 40.10, one of the effects of diffuse scattering is a shift between the delay of the specular reflection and the peak of the waveform (cross-correlation function between the signal and the receiver synthesized replica). Reflections off smooth surfaces keep the triangular shape of the autocorrelation function, and its peak corresponds to the specular ray path. Reflections off rough surfaces contain multiple ray paths, which delay the peak of the total waveform with respect to the shortest ray path (specular). The delay of the specular point can be obtained either by fitting a model to the waveform, or as the point of maximum slope within its leading edge. Once the delay of the specular reflection has been extracted, the altimetric observable is its pseudorange ρ_R in absolute terms or with respect to the direct pseudorange ρ_D : $\rho_R - \rho_D$

certain surface roughness status (more details will be given in Sect. 40.3.2 and (40.45)) or to estimate simultaneously both the specular delay and the roughness parameters (e.g., [40.31]). An alternative approach was suggested in [40.33] and later studied in [40.32], for which the delay of the specular ray path is identified as the point in the leading edge of the waveform of max-

imum slope (or equivalently, the peak of the derivative of the waveform, see Fig. 40.10).

Once the specular ray-path delay has been estimated through either fitting a model or looking at the derivative peak, this links to its pseudorange ρ_R , now given in units of length. The pseudorange can be given in absolute terms or with respect to the range of the line-

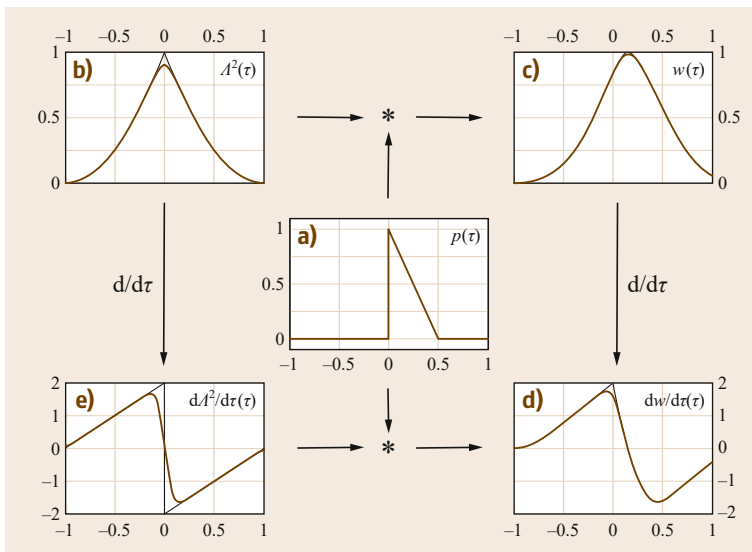


Fig. 40.10a-e The convolution between the power response of the sea surface when some power is scattered off nonspecular areas ((a) p for $\tau \geq 0$, where zero is defined as the specular ray-path delay) and the power of the autocorrelation function of the C/A modulation code ((b) Λ^2) results in a waveform ((c) w) the peak of which is delayed with respect to the specular delay. The derivative of this waveform presents a peak in the actual specular delay ((d) and (e) shows the derivative of the power of the autocorrelation). Note that this would strictly hold for a receiver with infinite bandwidth (thin black lines), while actual receiver filtering smooths these figures and slightly shifts the delays (thick brown lines) (after [40.32], courtesy of Institute of Electrical and Electronics Engineers (IEEE))

of-sight radio link, ρ_D . This is illustrated in Fig. 40.9. The pseudorange of the reflected signal has several contributions

$$\begin{aligned} \rho_R(t) = & \rho_{\text{geo}} \left(R(t), T \left(t - \frac{\rho_{\text{geo}}}{c} \right), S(R, T, h) \right) \\ & + \rho_{\text{iono}} + \rho_{\text{tropo}} + \rho_{\text{clk}} + \rho_{\text{ins}} + \epsilon. \end{aligned} \quad (40.42)$$

As in direct reception, the pseudorange has a geometric term, ρ_{geo} , given by the actual distance between the transmitter position T and the specular point position S , and the distance between the specular point and the receiver position R . The location of the specular point S is a function of the locations of the source and the receiver, as well as the altitude of the reflecting surface, h , with respect to a well-known reference surface (Earth geoid or ellipsoid). Other terms that contribute to the pseudorange are the tropospheric and ionospheric delays (ρ_{tropo} , ρ_{iono}) discussed in Chap. 6, clock drifts (ρ_{clk}), instrumental errors ρ_{ins} (such as those introduced by the antenna phase and cabling offsets), and, finally, other noise contributions ϵ (essentially thermal and speckle due to the random nature of the scattering surface). Note that the vertical component h of the specular point location is the unknown parameter in altimetric applications. An a priori value is assumed, based on a topographic description of the Earth's surface S evaluated at the point such that generates the shortest reflection link between the transmitter and the receiver. This point also verifies the Fresnel law of reflection (equal incident and scattering angles). Because the a priori knowledge of the Earth's topography wants to be refined, local vertical corrections h are expected, and solved for. The general form to solve h is to minimize

$$\min \left\{ |\rho_R(t)^{\text{obs}} - \rho_R(t)^{\text{mod}}(h)| \right\}. \quad (40.43)$$

The model of the pseudorange ρ_R^{mod} must include correction terms for its nongeometric components and/or estimate them together with the altimetric solution h .

This general term can be greatly simplified when the receiver is located at low altitude and the Earth is locally flat. Then the range of the reflected signal relative to the direct line-of-sight range is simply

$$\Delta \rho = \rho_{R,\text{geo}} - \rho_{D,\text{geo}} = 2H \sin(E), \quad (40.44)$$

where now H is the vertical distance between the surface and the receiver altitude, and E is the elevation angle of observation above the horizon (complementary to the incidence angle).

These techniques have been tested in a set of experimental campaigns, and their results published in

specialized journals, as cited below. Four of them present altimetric performance from aircrafts at 1–3 km altitude using the C/A clean-replica approach. All of them, conducted with different equipment and relatively low antenna gains, agree that the uncertainty of one-second integrated altimetric solutions from single-GNSS satellite signals is of the order of 1.5 m at near-nadir observation geometries [40.32, 34–36]. Another airborne experiment, for which P(Y) GPS codes were accessible, showed that this level of precision was slightly improved when using precise codes (≈ 1.4 m in one-second integration [40.31]). As can be deduced from Fig. 40.5, the range precision of the interferometric group-delay measurements should be better than the C/A clean-replica ones.

There are only two experimental campaigns in which the interferometric approach was tested, one from a bridge at 18 m altitude and over estuary waters [40.30], another from an aircraft flying at 3 km altitude over open sea waters [40.12, 36]. Over open waters the interferometric performance was at the 0.6 m level in one-second integration, but much better (~ 7 cm) over estuary waters. The factor of two between achieved altimetric precisions with the C/A clean-replica and the interferometric approaches was also predicted from a realistic stochastic model in [40.12], a study that indicates that this upper-bound factor might also occur from spaceborne receivers. A factor of three between both techniques was obtained from analytical models as shown in [40.37].

40.3.2 Sea Surface Scatterometry

As illustrated in Fig. 40.9, the roughness of the reflecting surface has a distorting effect in the received waveform. The analysis of this distortion thus permits the inference of the roughness properties of the surface. When the scattering surface is the ocean, the estimation of its roughness is linked to phenomena like surface winds and swell, of geophysical and civil interest. This is the aim of the sea surface scatterometric applications.

A model of the reflected waveform has been given in (40.25) to (40.28). The latter includes the bistatic scattering coefficient σ^0 , the only factor affected by the roughness. A widely accepted expression for σ^0 at each location P on the sea surface is [40.16]

$$\sigma^0(P) = \frac{\pi |R|^2 q^4}{q_z^4} \text{PDF} \left(-\frac{q_{\perp}}{q_z} \right), \quad (40.45)$$

where $|R|$ is the Fresnel coefficient, which in turn depends on the permittivity of the water. The vector \mathbf{q} is the scattering vector defined as

$$\mathbf{q} = k(\hat{\mathbf{u}}_{\text{scatt}} - \hat{\mathbf{u}}_{\text{inc}}),$$

where k is the electromagnetic wave number and $\hat{\mathbf{u}}$ denotes unit vectors along the scattered and incident directions. Furthermore, q_z and $\mathbf{q}_\perp = (q_x, q_y)^T$ are the vertical and horizontal components of the scattering vector. Note that those depend on the surface location P. The PDF is the probability density function of the two-dimensional (2-D) slopes $\mathbf{s} = (s_x, s_y)^T$ of the surface Σ . The particular slope given by $\mathbf{s} = -\mathbf{q}_\perp/q_z$ corresponds to the slope required at the point P of scattering vector \mathbf{q} to forward the signal to the receiver location (Fig. 40.11). The slopes' PDF is thus a statistical description of the surface roughness state.

The sea surface slopes' PDF is generally assumed as a Gaussian distribution, parametrized by the variance of the slopes or mean squared slopes (MSS). This first approach might be too restrictive, since it assumes isotropic surfaces. The next level of complexity, introduced to account for anisotropy in the roughness model, is given by a bivariate normal PDF of the slopes, parametrized by the variances of the slopes in orthogonal directions

$$\text{PDF}(\mathbf{s}) = \frac{1}{2\pi \sqrt{\text{Det}(\mathbf{M})}} e^{-\frac{1}{2} \mathbf{s}^T \mathbf{M}^{-1} \mathbf{s}} \quad (40.46)$$

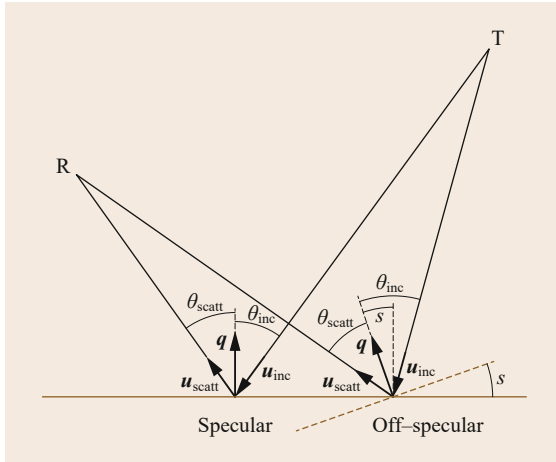


Fig. 40.11 In the optical geometry limit, the reflection Fresnel law must apply locally, that is, $\theta_{\text{inc}} = \theta_{\text{scatt}}$. When reflection occurs off the specular point, the surface must be tilted to accomplish it. The tilt, of slope s , is given by the ratio between the horizontal and the vertical components of the scattering vector $\mathbf{q} = k(\hat{\mathbf{u}}_{\text{scatt}} - \hat{\mathbf{u}}_{\text{inc}})$. When the surface is locally oriented in this way, the signal reflected off that point can reach the receiver. The probability that the surface has such local tilt depends on its roughness state, through the $\text{PDF}(\mathbf{s})$: the rougher the surface is, the higher is the probability of appropriate local tilts and hence the wider the glistening zone will be

with \mathbf{M} denoting the slopes' covariance matrix

$$\begin{aligned} \mathbf{M} &= \begin{pmatrix} \kappa_{20} & \kappa_{11} \\ \kappa_{11} & \kappa_{02} \end{pmatrix} \\ &= \mathbf{R}(\phi_w) \begin{pmatrix} \kappa_w & 0 \\ 0 & \kappa_c \end{pmatrix} \mathbf{R}(\phi_w)^{-1}. \end{aligned} \quad (40.47)$$

Here, κ_w and κ_c are variances along the wind direction and crosswind direction respectively, and $\mathbf{R}(\phi_w)$ is the rotation matrix by the angle ϕ_w defined between the upwind direction and the x -axis. The isotropic MSS links to these variances through $\text{MSS} = \kappa_w + \kappa_c$. More realistic slope distributions can be achieved by setting the PDF as a Gram–Charlier distribution [40.38] at the cost of a much larger number of parameters, which results in a much more complex inversion strategy.

The variances of the sea surface slopes can be obtained from the ocean wave spectrum $\Psi(\mathbf{k})$, given in the domain of the wave numbers \mathbf{k} , as

$$\begin{aligned} \text{MSS} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (k_x^2 + k_y^2) \Psi(\mathbf{k}) d\mathbf{k} \\ &= \text{MSS}_x + \text{MSS}_y. \end{aligned} \quad (40.48)$$

Some examples of accepted sea-surface wave spectra are Pierson–Moskowitz [40.39], JONSWAP [40.40], and the most widely used in GNSS-R is given by [40.41]. These are all wind-wave spectra, to which low frequency components can be added to account for swell. These are also used to link wind speed information with slopes' variances, and thus permit direct inversion of the wind parameters.

Extraction of the Roughness Parameters

The glistening zone is in general much larger than the Fresnel zone of coherent scattering. Glistening zones of a few hundred km extension for receivers in low Earth orbit (LEO) have been reported from theoretical studies (e.g., [40.42]) and checked from data of the UK-DMC mission [40.23]. As described in Sect. 40.2 the power gathered by the waveform comes from a certain area within the glistening zone, constrained by the ambiguity function $|\chi|^2$, which in turn depends on the delay of the correlation τ and the coherent integration time T_c . The latter imposes a filter in the frequency domain, of $1/T_c$ half-width, which splits the glistening zone into a set of frequency belts, the energy reflected from which will only be captured when the correlation is performed after shifting the central frequency to a frequency within each belt. This was illustrated in Fig. 40.4. Therefore, the waveform (also called delay map, DM) obtained at the frequency of the nominal specular point does not

capture the totality of the scattered energy, but only the signals reflected off the specular frequency belt. By varying the frequency at which the correlation is performed, it is then possible to obtain the delay maps over other belts. This corresponds to varying $\delta\nu$ in (40.25). The composition of all these delay maps, taken at different frequencies, is called a delay-Doppler map (DDM), and it is sketched in Fig. 40.12.

The first set of algorithms that were used to extract roughness information from the DM and/or DDM observables were based on fitting a model such as (40.25) to the data. Given the geometry of the observation and instrumental information such as antenna gains, the only unknown parameters are those related to the PDF of the surface slopes (variances or MSS, or their expressions as a function of the wind speed). Several experiments were conducted to perform sea surface scatterometric GNSS-R, in which the roughness was obtained by fitting a model to the data. In most of these cases the PDF of the slopes was assumed as either Gaussian or bivariate normal distributions. The retrieval using bivariate normal distributions aimed at obtaining anisotropic roughness information, such as wind speed and direction, or κ_w and κ_c . To avoid singularities in the anisotropic retrievals it was necessary to use either the full DDM information and/or to combine DM information from multiple satellites.

Other algorithms to extract roughness information from the reflected waveforms are compiled in Table 40.5, together with their achievements and bibliographic references.

40.3.3 Sea Surface Permittivity

GNSS signals transmit L-band signals, a portion of the microwave spectrum sensitive to variations in the temperature and salinity of the surface water (through variations in its permittivity). In fact, two spaceborne radiometric missions have chosen L-band measurements to sense sea surface salinity (SMOS [40.54] and Aquarius [40.55]). Little work has been done with GNSS reflections to infer permittivity parameters of the ocean. The studies found in the literature are based on polarimetric observations, comparing either the power of co- and cross-polar components after the reflections (polarimetric ratio [40.21]), or the phase of these components (polarimetric phase interferometry, POPI [40.56]). Both observables present small sensitivity, for instance $\sim 4\%$ variations in the polarimetric ratio (as deduced from the Fresnel coefficients) when changing the surface salinity from 25–40 pps. The effects of roughness need to be addressed, as well as a proper model of the phase wind-up in reflected signals for the POPI technique [40.57].

40.3.4 Cryosphere: Ice and Snow

Snow Depth Around Ground-Based GNSS Stations

Geodetic institutions keep networks of permanent ground-based GNSS stations for precise monitoring of the Earth's crust. Some of these stations are in regions covered by seasonal snow. The reflection of the GNSS signals off the snow interferes with the line-of-sight sig-

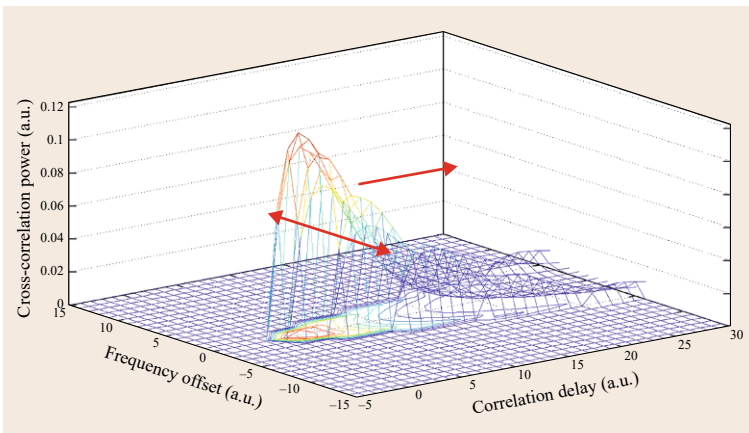


Fig. 40.12 Sketch of a delay-Doppler map for GNSS signals after a diffuse scattering process such as reflection off ocean waters; the cross-correlation function (its power given in (40.25)) evaluated at different delay τ and frequency $\delta\nu$ values. All units are arbitrary in this illustration. As the roughness increases, the peak power decreases, and the energy spreads both along the delay and the frequency axis. This distortion is used to infer roughness information, for example through the slope of the trailing edge, or fitting theoretical models, or quantifying its volume and area (see Table 40.5 for other observables and inversion algorithms)

Table 40.5 Summary of some GNSS-R experiments and their techniques for sea surface scatterometric applications

Description	Achievements	Refs.
Model fitting		
Fitting data from a single GNSS satellite to a waveform model that was generated with a slope's PDF parametrized directly with the wind speed parameter or the slope variances	2 m/s agreement in wind speed retrieval up to 9 m/s wind speeds	[40.43, 44]
Same procedure but estimating wind direction ($\pm 180^\circ$ ambiguity) by joint inversion of multiple simultaneous GNSS signals	2 m/s wind speed and 30° wind direction agreement with a wind scatterometer, up to 10 m/s wind speeds	[40.45, 46]
Heuristic correction term in the relationship between MSS and wind u (in m/s) to extend its validity to high wind speeds (hurricanes) $\text{MSS}_w(u) = 0.45(0.00316f(u))$ $\text{MSS}_c(u) = 0.45(0.003 + 0.00192f(u))$ <p>with</p> $f(u) = \begin{cases} u & (0 < u \leq 3.49) \\ 6 \ln(u) - 4 & (3.49 < u \leq 46) \\ 0.411u & (46 < u) \end{cases}$	Wind speeds of up to over 40 m/s in hurricane conditions measured without saturation, ≈ 5 m/s root mean square (RMS) accuracy compared to radiosonde measurements at high wind speeds.	[40.47, 48]
Fitting a model to the entire DDM	To extract anisotropic MSS from each single GNSS satellite	[40.49]
Fitting a model to the trailing edge of the DM solely	Wind speed with precisions better than 2 m/s, with winds up to 10 m/s	[40.16, 50]
Integrated spectra		
Integration of the DDM along either the frequency axis to obtain the integrated delay function $\text{IDM}(\tau) = \int_{\text{Dopplers}} \text{DDM}(\tau, f) df$ <p>or along the delay axis to obtain the integrated Doppler spectrum</p> $\text{IDS}(f) = \int_{\text{Delays}} \text{DDM}(\tau, f) d\tau$ <p>The width or spread of these functions relates to roughness parameters: the frequency width</p> $B_{3\text{dB}} = 4\sqrt{2 \ln 2} \sin E / \lambda \sqrt{v_s^T \mathbf{M} v_s + \epsilon}$ <p>being \mathbf{M} the covariance matrix of the surface slopes, and v_s the velocity of the specular point with respect to the receiver; and</p> $\text{Var}(\tau) \propto \left(\kappa_{20}^2 + 2\kappa_{11}^2 \sin^{-2} E + \kappa_{02}^2 \sin^{-4} E \right)$ <p>being κ elements of \mathbf{M}</p>	Theoretical study supported by experimental airborne data and later applied to UK-DMC LEO data	[40.23, 51]

nals to generate interference patterns in the received data. These features were first used to extract the snow depth around these ground stations [40.58]. The observable analyzed to infer the snow depth is the frequency of the multipath-induced oscillations in the SNR, as pic-

tured in Fig. 40.13. The technique has been applied to the USA UNAVCO's Plate Boundary Observatory (PBO) network, and resulting products made available to the public at [40.59] and documented in [40.60]. The technique has been improved in [40.61, 62]. Over-

Table 40.5 (continued)

Description	Achievements	Refs.
Scatterometric delay The peak of the DM is delayed with respect to the specular ray path, and this effect depends on the MSS.	Near-real-time MSS from airborne campaigns at 0.001 MSS precision level	[40.52]
DDM area and volume Direct link between the area or volume of the DDM (above a certain threshold) and the roughness state of the surface.	Experimental work done in the frame of SMOS L-band roughness corrections.	[40.19]
Discrete PDF Rewriting the radar equation of the DDM to express it as a series of terms evaluated in cells of similar required slope s_{ij} (slope needed for forwarding the signal towards the direction of the receiver). This series is linear with respect to the PDF of the slopes, which can thus be linearly inverted in the set of discrete values s_{ij} $W(\tau, f) = \sum_{s_{ij}} A_{s_{ij}}(\tau, f) \text{PDF}(s_{ij}) .$	Data from airborne experiments have been inverted to discrete slopes' PDF. The results present robustness to selection of the discrete sampling; consistency with MSS values obtained from retrievals based on Gaussian and bivariate normal distributions; agreement with independent wind measurements at 2 m/s level; it sensed non-Gaussian features such as the sense of the wind, consistent with independent measurements	[40.53]

all, the methodology performs within the limitations of both GPS and in situ measurement errors, resulting in a correlation of 0.98 and RMS error of 6–8 cm for observed snow depths of up to 2.5 m, with the GPS underestimating in situ snow depth by 10–15%. A similar project is being implemented in French geodetic networks [40.63].

A similar approach, but measured with a linear V-polarized antenna, was suggested in [40.64]. In that case, the information is obtained from the notches of the interference pattern (location and amplitude). Both techniques have also been used to extract other geophysical information such as soil moisture and vegetation cover (see Sect. 40.3.5).

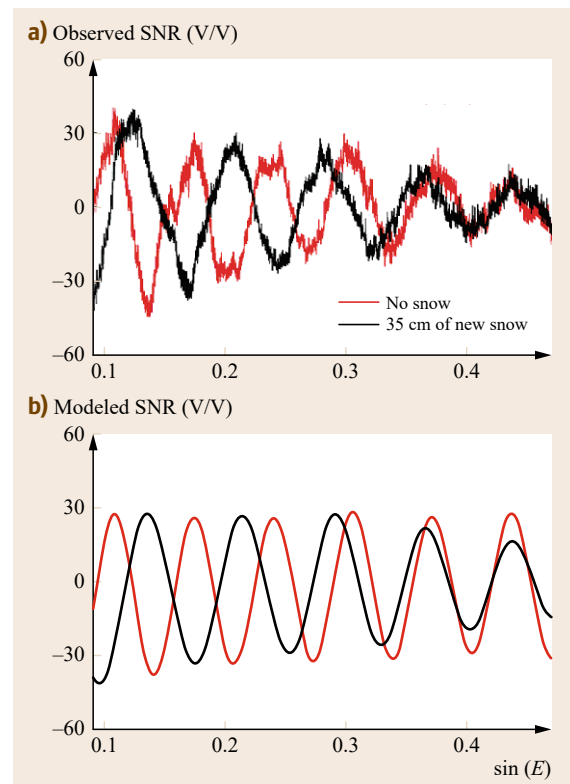
Sea Ice

Early experimental studies on the potential use of GNSS reflectometry for sea ice monitoring, conducted aboard an aircraft over Arctic sea ice [40.65], have shown that the peak power changed significantly along the track, with positive correlation with the backscattering measurements of the RADARSAT installed in the same aircraft. This positive correlation between forward and backward scattering was an indication of variations in the dielectric properties of the ice rather than roughness (roughness effects produce negative

correlation between forward and backward scattering powers).

Sea ice roughness was tackled in [40.66], which evolved a mixed technique to separate roughness parameters from permittivity estimates of the sea ice: after renormalization of the reflected waveform, its peak

Fig. 40.13 (a) Patterns captured in GNSS SNR observables in a geodetic station due to the interference between the line-of-sight signal and the signal reflected off the surrounding snow. (b) Modeled interference patterns. Both graphs show the variation of the signal-to-noise ratio (SNR) with the sine of the elevation angle E (after [40.58], courtesy of Wiley) ►



power is mapped to an absolute ice skin permittivity via the cross-polar Fresnel coefficients, being the skin permittivity the effective permittivity of the ice at its upper layer, as thick as the penetration depth of the L-band signals. On the other hand, the inversion of the roughness parameter, in the form of MSS, was based on fitting a model to the shape of the renormalized waveform.

Polarimetric observations have also been proven to relate to the sea ice permittivity state. Figure 40.14 shows the polarimetric ratio defined as the copolar power divided by the cross-polar one, from an experiment on top of a cliff ≈ 700 m high, overlooking Disko Bay in Greenland. The geometry of the experiment forced the elevation angles to be around the Brewster angle, which in circular polarization separates reflectivity of the higher circular cross-polar component (above the Brewster elevation) from reflectivity of the higher circular copolar component (below the Brewster elevation).

Finally, another aspect of some types of sea ice is its relative smoothness compared to open waters. When this happens, the reflection has a strong coherent component, which enables the connection of the phase. Then, phase-delay measurements are possible as shown in [40.68, 69], even from space-based observations when in grazing geometries [40.70]. In nadir-like geometries from spaceborne platforms, phase-delay measurements have been proven to be more difficult, at

least using the only two sea-ice reflection observations captured with the UK-DMC satellite [40.25].

In a similar way, glacial lake outburst floods have been monitored using GNSS-R techniques [40.71].

Dry Snow

The first theoretical studies on the potential use of GNSS reflectometry to characterize the surface and interior of ice sheets was presented in [40.72]. The work derived a complex model to account for internal reflections and volumetric scattering, and conducted simulations to show that the reflected signals are sensitive to firm parameters such as snow accumulation rate.

The actual penetration of the GNSS signals into Antarctic dry snow was experimentally proven in [40.67, 73], with data obtained from a GNSS-R instrument installed at 45 m altitude, on the American Tower of Concordia Station. The study explained the unusual behavior of the GNSS-R data by means of a model that accounted for multiple coherent reflections occurring at several internal layers of the snow structure. These reflections were captured down to approximately 300 m depth. The observable used for this application is the lag hologram, inspired by the radio holograms developed in other fields such as in GNSS Radio occultations (Chap. 38).

Lag holograms present the spectral content of the received signal with respect to a reference signal (here

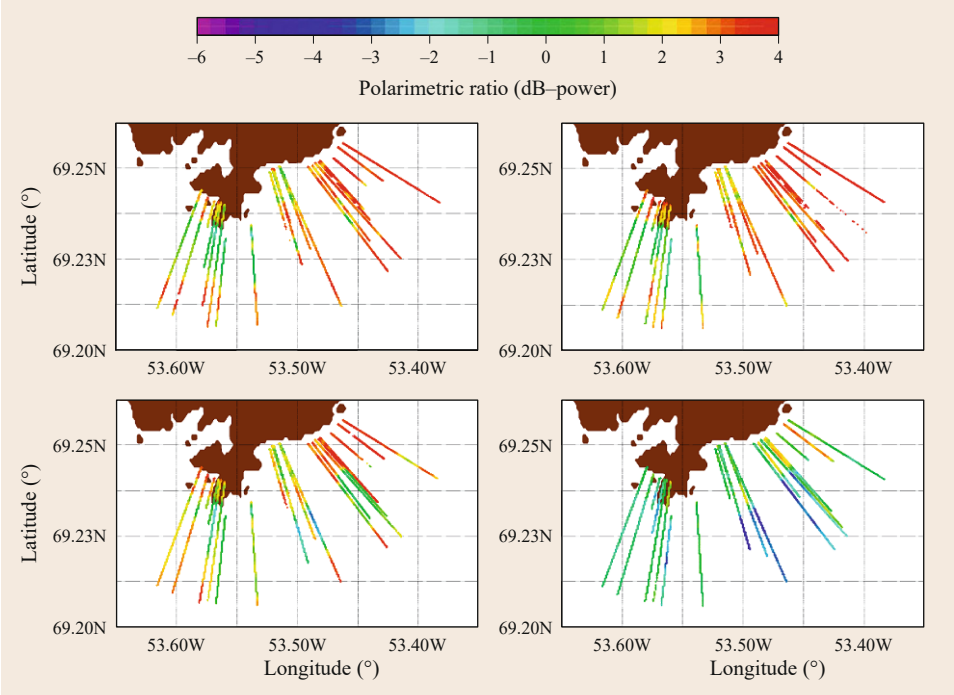


Fig. 40.14 Polarimetric ratio of GNSS-R observations over Disko Bay, Greenland, taken from ≈ 700 m high cliff. It corresponds to four sequential days, day number 44 of year 2009 to day number 47. The temperature dropped during days 46 and 47 (bottom) (after [40.67])

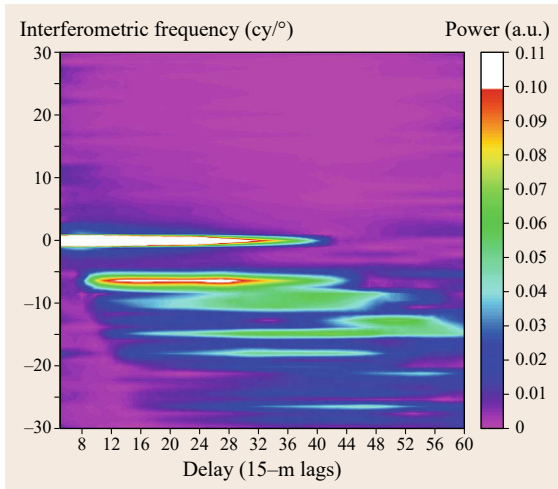


Fig. 40.15 Example of a lag hologram extracted from a time series of 128 one-second complex waveforms in a GNSS-R dry snow experiment at Concordia Station (Antarctica). It corresponds to pseudo-random noise (PRN) 13, 16 December 2009, between 44.5° and 45.5° elevation. The x -axis is the correlation delay τ , given in lags (15 meter interlag distance), and with respect to an arbitrary zero (being the delay of the direct signal at around lag number 20 and the delay of the surface reflection at around lag number 24). The frequency is given as interferometric cycles per degree elevation, which relates to the depth of the reflecting layer. The zero frequency corresponds to the reference field: the line-of-sight ray. The geometry of the observation results in frequencies more negatives than -5.8 cy/deg corresponding to scattering off reflecting elements below the snow-air interface. Note different spectral content for different delay lags along the waveform (after [40.73], courtesy of Elsevier Ltd.)

the direct line-of-sight one), for each of the waveform's correlation delays (lags). If the spectral information were obtained from the peak of the waveform solely (as it is done in radio occultation radio holograms), it would be limited by signals arriving up to one C/A-code chip delayed (~ 300 m). Because internal reflections can occur down to 300 m, their delay with respect to the peak can be as long as twice this range, and thus filtered out by the code modulation. Inspecting the spectral content of the entire waveform permits the extraction of information on deeper layers (see, e.g., the example of an Antarctic lag hologram in Fig. 40.15).

40.3.5 Land: Soil Moisture and Vegetation

For several reasons, the L-band, between frequencies of 1 and 2 GHz, is one of the spectral bands that work bet-

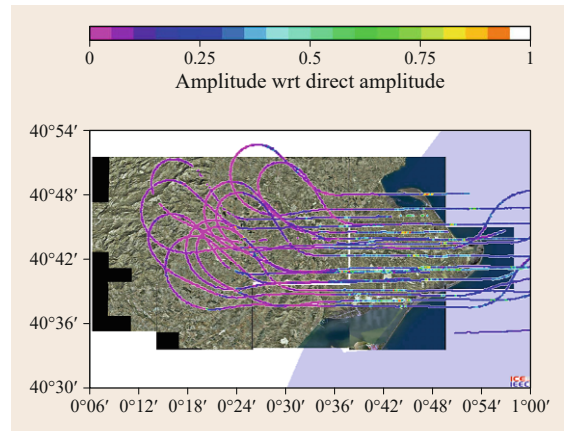


Fig. 40.16 Flight over Ebro River delta with the GOLD-RTR GNSS-R dedicated receiver [40.52]: the amplitude of the reflected signals, normalized by the amplitude of the direct line-of-sight signals, is plotted over an approximate mosaic of aerial photographs. The delta of the river, covered by wetlands and rice crops, contrasts with the dry surrounding mountains, as captured by the GNSS-R signals

ter for soil moisture monitoring. At these frequencies the atmosphere is effectively transparent and vegetation is semitransparent, the microwave measurement is strongly dependent on soil moisture, and the measurements are independent of solar illumination. The GNSS signals are sensitive to the moisture content in the upper layers (1–6 cm depth; see [40.74]).

The first algorithms to extract soil moisture information from GNSS reflectometry data were based on the peak power or amplitude of the reflected waveform, often renormalized as a way to calibrate geometric and instrumental issues. Examples of these early approaches can be found in for example [40.76, 77], and in the data shown in Fig. 40.16.

Despite vegetation being semitransparent to L-band signals, it does affect the reflection process, together with the surface roughness. Untangling the different components of the process requires complex modeling, which involves both the coherent and incoherent surface scattering, together with volumetric scattering through the vegetation structures. To help in inferring these components, either polarimetric (e.g., [40.78]) or interferometric (e.g., [40.75, 79, 80]) approaches have been followed. Figure 40.17 shows the theoretical sensitivity of the interferometric pattern to soil and vegetation parameters, as captured with a vertical-polarization antenna. This technique uses the location and amplitude of the interference notches, as it is also used for retrieval of snow depth (Sect. 40.3.4).

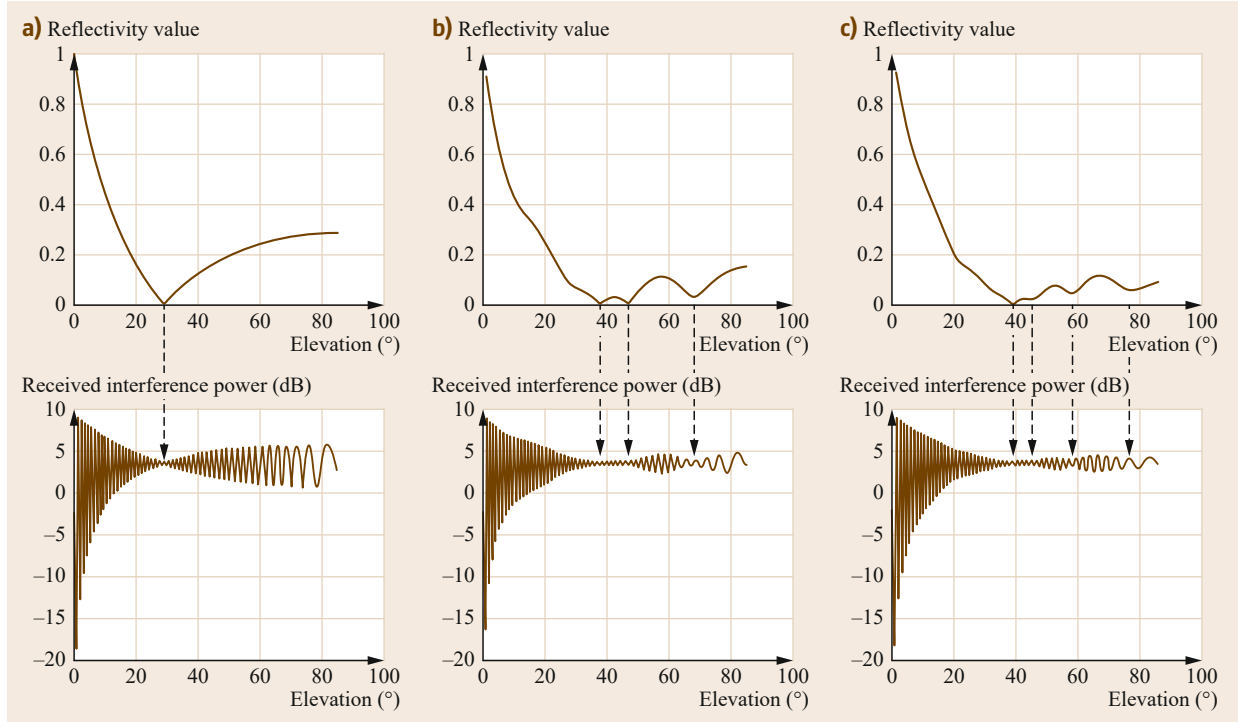


Fig. 40.17a–c Simulated interference power received versus reflectivity for three cases of vegetation-covered soils. **(a)** Bare soil produces one notch, **(b)** 60 cm vegetation layer + soil layer produces three notches, and **(c)** 90 cm vegetation layer + soil layer produces four notches (after [40.75], courtesy of IEEE)

40.4 Spaceborne Missions

Several GNSS-R experiments have been conducted from spaceborne platforms:

Space Shuttle/SIR-C The first GNSS reflection captured from the outer Space was collected with the SIR-C instrument aboard the Shuttle. The study, presented in [40.22], was conducted in postprocessing, searching for GPS signals in archived datasets obtained with the Space Shuttle imaging radar system. This achievement proved the possibility of receiving GNSS signals reflected off the sea surface from space.

CHAMP was a German geoscience satellite operating from 2000 to 2010, which included a payload for GNSS radio occultation (RO) observations (see Chap. 38 for details about this technique). *Beyerle et al.* [40.81] first showed that CHAMP RO data contained embedded GNSS signals reflected off the Earth's surface. *Cardellach et al.* [40.70] used GNSS reflected signals collected from the RO payload of CHAMP to perform phase-delay altimetry in polar regions.

UK Disaster Monitoring Constellation (UK-DMC)

was a British satellite built by Surrey Satellite Technology Ltd. (SSTL) that included a dedicated GNSS-R payload for experimental purposes. The mission was active between 2003 and 2011. A limited set of GNSS-R data streams were acquired from ≈ 700 km altitude with a moderate antenna gain (11.8 dBi) and postprocessed on the ground. The results, published in [40.23] to [40.25], proved the sensitivity of spaceborne GNSS-R to sea surface roughness and sea ice, and the possibility of receiving land reflections from space. Sets of GNSS-R data obtained with the UK-DMC Low Earth Orbiter are also available at [40.26].

UK TechDemoSat-1 (UK-TDS1)

is a microsatellite (of about 50 kg payload) built by Surrey Satellite Technology Ltd. (SSTL). Among other payloads, it carries a new SSTL GNSS-R receiver (a clean-replica approach) for ocean roughness applications among other payloads. UK-TDS1 was successfully launched in July 2014, and since then it has been acquiring GNSS-R in duty cycles of two days out

of eight. Initial results from UK-TDS1 data can be found in [40.82].

After years of research, feasibility studies and technology development, the GNSS-R techniques are entering a maturity phase. Besides UK-TDS1, currently orbiting, two other spaceborne GNSS-R missions were launched in 2016. The missions use nano- to microsatellites, and the GNSS-R receiver technique will be of the clean-replica type. They are mostly focused on scatterometric applications:

³CAT-2 is a six-unit CubeSat (<10 kg) built by the Politechnical University of Catalonia (UPC) that will carry UPC's PICARO receiver. The receiver uses a clean-replica approach, but will also attempt P-code retrievals [40.83]. It was launched in August 2016.

CYGNSS is a constellation of eight satellites, with three-unit CubeSat each, funded by the National Aeronautics and Space Administration (NASA) Venture program and led by the University of Michigan [40.84]. They will carry the new SSTL GNSS-R receiver tested in UK-TDS. The constellation will be allocated in low inclination orbits to monitor tropical storms. The mission was launched in December 2016.

In addition to the launch of these missions, another mission of the European Space Agency (ESA) is going through the initial stages of development (feasibility

phase-A studies, technology breadboarding and prototyping, etc.). Its final launch and operations are not guaranteed at this stage of development. The main difference with respect to the other GNSS-R missions is the scientific objective, mostly focused on altimetric applications:

GEROS-ISS (GNSS rEfectometry Radio Occultation and Scatterometry on board the International Space Station) is an ESA experiment aboard the International Space Station (ISS), in low inclination orbit, the main objective of which is to perform mesoscale altimetry with GNSS-R without some of the physical constraints of the small satellite platforms [40.85]. The mission requirements and system requirements documents have been issued and two phase-A feasibility studies are being executed.

All of these missions will help to further advance the state-of-the art in spaceborne reflectometry and scatterometry, an emerging GNSS application that promises substantial scientific and social benefits.

Acknowledgments. The authors would like to acknowledge the support of the Spanish grant AGORA: Advanced GNSS and other signals of Opportunity Reflectometry for Accurate Climate Monitoring (ESP2015-70014-C2-R). The authors participate in the EUMETSAT ROM SAF.

References

- 40.1 M. Martín-Neira: A passive reflectometry and interferometry system (PARIS): Application to ocean altimetry, *ESA Journal* **17**, 331–355 (1993)
- 40.2 J.L. Garrison, S.J. Katzberg, M.I. Hill: Effect of sea roughness on bistatically scattered range coded signals from the Global Positioning System, *Geophys. Res. Lett.* **25**(13), 2257–2260 (1998)
- 40.3 S. Jin, E. Cardellach, F. Xie: *GNSS Remote Sensing* (Springer, Dordrecht 2014)
- 40.4 M. Born, E. Wolf: *Principles of Optics*, 7th edn. (Cambridge Univ. Press, Cambridge 1999)
- 40.5 A.R. Thompson, J.M. Moran, G.W. Swenson: *Interferometry and Synthesis in Radio Astronomy*, 2nd edn. (Wiley-VCH, New York 2001)
- 40.6 B.W. Parkinson, J.J. Spilker: *Global Positioning System: Theory and Applications* (AIAA, Washington 1996)
- 40.7 P.F. MacDoran: Satellite emission radio interferometric Earth surveying series – GPS geodetic system, *Bull. Géod.* **53**(2), 117–137 (1979)
- 40.8 J.D. Romney: Theory of correlation in VLBI. In: *Very Long Baseline Interferometry and the VLBA*, ed. by J.A. Zensus, P.J. Diamond, P.J. Napier (Astronomical Society of the Pacific, San Francisco 1995) pp. 17–35
- 40.9 R. Bracewell: *The Fourier Transform and Its Applications* (McGraw-Hill, New York 1965)
- 40.10 A.R. Whitney, R. Cappallo, W. Aldrich, B. Anderson, A. Bos, J. Casse, J. Goodman, S. Parsley, S. Pogrebenko, R. Schilizzi, D. Smythe: Mark 4 VLBI correlator: Architecture and algorithms, *Radio Sci.* **39**(RS1007), 1–24 (2004)
- 40.11 K. S. Andrews, A. Argueta, N. E. Lay, M. Lyubarev, E. H. Sigman, M. Srinivasan, A. Tkachenko: *Reconfigurable Wideband Ground Receiver Hardware Description and Laboratory Performance*, IPN Progress Report 42–180 (Jet Propulsion Laboratory, Pasadena 2010)
- 40.12 E. Cardellach, A. Rius, M. Martín-Neira, F. Fabra, O. Nogués-Correig, S. Ribó, J. Kainulainen, A. Camps, S.D. Addio: Consolidating the precision of interferometric GNSS-R ocean altimetry using airborne experimental data, *IEEE Trans. Geosci. Remote Sens.* **52**(8), 4992–5004 (2014)

- 40.13 P.Z. Peebles: *Radar Principles* (Wiley, Hoboken 2004)
- 40.14 B.C. Barker, J.W. Betz, J.E. Clark, J.T. Correia, J.T. Gillis, S. Lazar, K.A. Rehborn, J.R. Straton: Overview of the GPS M code signal, *Proc. ION NTM 2000*, Anaheim (ION, Virginia 2000) pp. 542–549
- 40.15 F.T. Ulaby, R.K. Moore, A.K. Fung: *Microwave Remote Sensing: Active and Passive, Vol. II: Radar Remote Sensing and Surface Scattering and Emission Theory* (Addison-Wesley, Norwood 1982)
- 40.16 V.U. Zavorotny, A.G. Voronovich: Scattering of GPS signals from the ocean with wind remote sensing application, *IEEE Geosci. Remote. Sens.* **38**(2), 951–964 (2000)
- 40.17 M. Martín-Neira, S. D'Addio, C. Buck, N. Floury, R. Prieto-Cerdeira: The PARIS ocean altimeter in-orbit demonstrator, *IEEE Geosci. Remote. Sens.* **49**(6), 2209–2237 (2011)
- 40.18 C. Zuffada, T. Elfouhaily, S. Lowe: Sensitivity analysis of wind vector measurements from ocean reflected GPS signals, *Remote Sens. Environ.* **88**(3), 341–350 (2003)
- 40.19 E. Valencia, A. Camps, J.F. Marchan-Hernandez, N. Rodriguez-Alvarez, I. Ramos-Perez, X. Bosch-Lluis: Experimental determination of the sea correlation time using GNSS-R coherent data, *IEEE Geosci. Remote Sens. Lett.* **7**(4), 675–679 (2010)
- 40.20 GOLD_RTR MINING web server for GNSS-R experimental data and related information (Institut de Ciències de l'Espai) http://www.ice.csic.es/research/gold_rtr_mining
- 40.21 E. Cardellach, F. Fabra, O. Nogués-Correig, S. Oliveras, S. Ribó, A. Rius: GNSS-R ground based and airborne campaigns for ocean, land, ice, and snow techniques: Application to the GOLD-RTR data sets, *Radio Sci.* **46**(RS0C04), 1–16 (2011)
- 40.22 S. T. Lowe, J. L. LaBrecque, C. Zuffada, L. J. Romans, L. E. Young, G. A. Hajj: First spaceborne observation of an Earth-reflected GPS signal, *Radio Sci.* **37**(1), 7.1–7.28 (2002)
- 40.23 S. Gleason, S. Hodgart, Y. Sun, C. Gommenginger, S. Mackin, M. Adjrad, M. Unwin: Detection and processing of bistatically reflected GPS signals from low Earth orbit for the purpose of ocean remote sensing, *IEEE Trans. Geosci. Remote Sens.* **43**(6), 1229–1241 (2005)
- 40.24 S. Gleason: Remote Sensing of Ocean, Ice and Land Surfaces Using Bistatically Scattered GNSS Signals from Low Earth Orbit, Ph.D. Thesis (University of Surrey, Surrey 2006)
- 40.25 S. Gleason: Towards sea ice remote sensing with space detected GPS signals: Demonstration of technical feasibility and initial consistency check using low resolution sea ice information, *Remote Sens.* **2**(8), 2017–2039 (2010)
- 40.26 S. Gleason, S. Lowe, V. Zavorotny: Remote sensing using bistatic GNSS reflections. In: *GNSS Applications and Methods*, ed. by S. Gleason, D. Gebre-Egziabher (Artech House, Boston 2009) pp. 399–436
- 40.27 R. Stosius, G. Beyerle, A. Helm, A. Hoechner, J. Wickert: Simulation of space-borne tsunami detection using GNSS-reflectometry applied to tsunamis in the Indian Ocean, *Nat. Hazards Earth Syst. Sci.* **10**, 1359–1372 (2010)
- 40.28 P.Y. Le Traon, G. Dibarboure, G. Ruffini, E. Cardellach: Mesoscale ocean altimetry requirements and impact of GPS-R measurements for ocean mesoscale circulation mapping, Technical Note Extract from the PARIS-BETA ESTEC/ESA Study, eprint <https://arxiv.org/abs/physics/0212068> (Starlab 2002)
- 40.29 M. Martín-Neira, M. Caparrini, J. Font-Rosselló, S. Lannelongue, C.S. Vallmitjana: The PARIS concept: An experimental demonstration of sea surface altimetry using GPS reflected signals, *IEEE Trans. Geosci. Remote Sens.* **39**(1), 142–149 (2001)
- 40.30 A. Rius, O. Nogués-Correig, S. Ribó, E. Cardellach, S. Oliveras, E. Valencia, H. Park, J.M. Tarongí, A. Camps, H. van der Marel, R. van Bree, B. Altena, M. Martín-Neira: Altimetry with GNSS-R interferometry: First proof of concept experiment, *GPS Solutions* **16**(2), 231–241 (2012)
- 40.31 S.T. Lowe, C. Zuffada, Y. Chao, P. Kroger, L.E. Young, J.L. LaBrecque: 5-cm precision aircraft ocean altimetry using GPS reflections, *Geophys. Res. Lett.* **29**(10), 13.1–13.4 (2002)
- 40.32 A. Rius, E. Cardellach, M. Martín-Neira: Altimetric analysis of the sea surface GPS reflected signals, *IEEE Trans. Geosci. Remote Sens.* **48**(4), 2119–2127 (2010)
- 40.33 G. Hajj, C. Zuffada: Theoretical description of a bistatic system for ocean altimetry using the GPS signal, *Radio Sci.* **38**(5), 10.1–10.19 (2003)
- 40.34 S.T. Lowe, C. Zuffada, J. LaBrecque, M. Lough, J. Lerma: An aircraft ocean altimetry measurement using reflected GPS signals, *Proc. IEEE IGARSS, Honolulu*, ed. by T.I. Stein (2000) pp. 1–3
- 40.35 G. Ruffini, F. Soulat, M. Caparrini, O. Germain, M. Martín-Neira: The eddy experiment: Accurate GNSS-R Ocean altimetry from low altitude aircraft, *Geophys. Res. Lett.* **31**(L12306), 1–4 (2004)
- 40.36 A. Rius, F. Fabra, S. Ribó, J.C. Arco Fernandez, S. Oliveras, E. Cardellach, A. Camps, O. Nogués-Correig, J. Kainulainen, E. Rohue, M. Martín-Neira: PARIS interferometric technique proof of concept: Sea surface altimetry measurements, *Proc. IEEE IGARSS, Munich* (2012) pp. 7067–7070
- 40.37 S. D'Addio, M. Martín-Neira: Comparison of processing techniques for remote sensing of earth-exploding reflected radio-navigation signals, *Electron. Lett.* **49**(4), 292–293 (2013)
- 40.38 C. Cox, W. Munk: Measurements of the roughness of the sea surface from photographs of the Sun's glitter, *J. Opt. Soc. Am.* **44**(11), 838–850 (1954)
- 40.39 W.J. Pierson, L. Moskowitz: A proposed spectral form for fully developed wind seas based on the similarity theory of A. A. Kitaigorodskii, *J. Geophys. Res.* **69**(24), 5181–5190 (1964)
- 40.40 K. Hasselmann, T.P. Barnett, E. Bouws, H. Carlson, D.E. Cartwright, K. Enke, J.A. Ewing, H. Gienapp, D.E. Hasselmann, P. Kruseman, A. Meerburg, P. Miller, D.J. Olbers, K. Richter, W. Sell, H. Walden: Measurements of wind-wave growth and swell decay during the Joint North Sea Wave Project (JONSWAP), *Dtsch. Hydrogr. Z.* **A8**(12), 95 (1973)

- 40.41 T. Elfouhaily, B. Chapron, K. Katsaros, D. Vandemark: A unified directional spectrum for long and short wind-driven waves, *J. Geophys. Res.* **102**(C7), 15781–15796 (1997)
- 40.42 E. Cardellach: Sea Surface Determination Using GNSS Reflected Signals, Ph.D. Thesis (Universitat Politècnica de Catalunya, Barcelona 2002)
- 40.43 A. Komjathy, V. Zavorotny, P. Axelrad, G. Born, J. Garrison: GPS signal scattering from sea surface: Wind speed retrieval using experimental data and theoretical model, *Remote Sens. Environ.* **73**(2), 162–174 (2000)
- 40.44 E. Cardellach, G. Ruffini, D. Pino, A. Rius, A. Komjathy, J.L. Garrison: Mediterranean balloon experiment: Ocean wind speed sensing from the stratosphere using GPS reflections, *Remote Sens. Environ.* **88**(3), 351–362 (2003)
- 40.45 A. Komjathy, M. Armatys, D. Masters, P. Axelrad: Retrieval of ocean surface wind speed and wind direction using reflected GPS signals, *J. Atmos. Ocean. Technol.* **21**(3), 515–526 (2004)
- 40.46 M. Armatys: Estimation of Sea Surface Winds Using Reflected GPS Signals, Ph.D. Thesis (University of Colorado, Boulder 2001)
- 40.47 S.J. Katzberg, O. Torres, G. Ganoe: Calibration of reflected GPS for tropical storm wind speed retrievals, *Geophys. Res. Lett.* **33**(L18602), 1–5 (2006)
- 40.48 S.J. Katzberg, J. Dunion, G.G. Ganoe: The use of reflected GPS signals to retrieve ocean surface wind speeds in tropical cyclones, *Radio Sci.* **48**(4), 371–387 (2013)
- 40.49 O. Germain, G. Ruffini, F. Soulat, M. Caparrini, B. Chapron, P. Silvestrin: The eddy experiment: GNSS-R speculometry for directional sea-roughness retrieval from low-altitude aircraft, *Geophys. Res. Lett.* **31**(L21307), 1–4 (2004)
- 40.50 J.L. Garrison, A. Komjathy, V.U. Zavorotny, S.J. Katzberg: Wind speed measurement using forward scattered GPS signals, *IEEE Geosci. Remote. Sens.* **40**(1), 50–65 (2002)
- 40.51 T. Elfouhaily, D.R. Thompson, L. Linstrom: Delay-Doppler analysis of bistatically reflected signals from the ocean surface: Theory and application, *IEEE Trans. Geosci. Remote Sens.* **40**(3), 560–573 (2002)
- 40.52 O. Nogués-Correig, E. Cardellach Gali, J. Sanz Campderros, A. Rius: A GPS-reflections receiver that computes Doppler/delay maps in real time, *IEEE Geosci. Remote. Sens.* **45**(1), 156–174 (2007)
- 40.53 E. Cardellach, A. Rius: A new technique to sense non-Gaussian features of the sea surface from L-band bi-static GNSS reflections, *Remote Sens. Environ.* **112**(6), 2927–2937 (2008)
- 40.54 Y. Kerr, J. Font, P. Waldteufel, M. Berger: The second of ESA's opportunity missions: The soil moisture and ocean salinity mission – SMOS, *ESA Earth Obs. Q.* **66**, 18–25 (2000)
- 40.55 D.M. Le Vine, F. Pellerano, G.S.E. Lagerloef, S. Yueh, R. Colomb: Aquarius: A mission to monitor sea surface salinity from space, *Proc. IEEE MicroRad, SanJuan* (2006) pp. 87–90
- 40.56 E. Cardellach, S. Ribó, A. Rius: Technical Note on Polarimetric Phase Interferometry (POPI), eprint <https://arxiv.org/abs/physics/0606099> (IEEC-CSIC 2006)
- 40.57 G. Beyerle: Carrier phase wind-up in GPS reflectometry, *GPS Solutions* **13**(3), 191–198 (2009)
- 40.58 K.M. Larson, E. Gutmann, V. Zavorotny, J. Braun, M. Williams, F. Nievinski: Can we measure snow depth with GPS receivers?, *Geophys. Res. Lett.* **36**(L17502), 1–5 (2009)
- 40.59 GPS Reflections Research Group: PBO H₂O Data Portal – Using GPS reflection data from NSF's Plate Boundary Observatory (PBO) to study the water cycle (Univ. of Colorado). <http://xenon.colorado.edu/portal>
- 40.60 K.M. Larson, F.G. Nievinski: GPS snow sensing: Results from the EarthScope Plate Boundary Observatory, *GPS Solutions* **17**(1), 41–52 (2013)
- 40.61 F.G. Nievinski, K.M. Larson: Inverse modeling of GPS multipath for snow depth estimation – Part I: Formulation and simulations, *IEEE Trans. Geosci. Remote Sens.* **52**(10), 6555–6563 (2014)
- 40.62 F.G. Nievinski, K.M. Larson: Inverse modeling of GPS multipath for snow depth estimation – Part II: Application and validation, *IEEE Trans. Geosci. Remote Sens.* **52**(10), 6564–6573 (2014)
- 40.63 K. Boniface, J. Braun, J. McCreight, S. Morin, F.G. Nievinski, A. Walpersdorf: GNSS interferometric reflectometry for snow depth measurements: Comparison to SNODAS model in the western US and first results in the French Alps, *Proc. Space Reflecto 2013, Brest* (Univ. du Littoral, Côte d'Opale 2013) pp. 1–2
- 40.64 N. Rodriguez-Alvarez, A. Aguasca, E. Valencia, X. Bosch-Lluis, A. Camps, I. Ramos-Perez, H. Park, M. Vall-Ilosera: Snow thickness monitoring using GNSS measurements, *IEEE Geosci. Remote Sens. Lett.* **9**(6), 1109–1113 (2012)
- 40.65 A. Komjathy, J. Maslanik, V.U. Zavorotny, P. Axelrad, S.J. Katzberg: Sea ice remote sensing using surface reflected GPS signals, *Proc. IEEE IGARSS, Honolulu*, ed. by T.I. Stein (2000) pp. 2855–2857
- 40.66 M. Belmonte Rivas, J.A. Maslanik, P. Axelrad: Bistatic scattering of GPS signals off arctic sea ice, *IEEE Trans. Geosci. Remote Sens.* **48**(3), 1548–1553 (2010)
- 40.67 F. Fabra: GNSS-R as a Source of Opportunity for Remote Sensing of the Cryosphere, Ph.D. Thesis (Universitat Politècnica de Catalunya, Barcelona 2013)
- 40.68 M. Semmling, G. Beyerle, R. Stosius, G. Dick, J. Wickert, F. Fabra, E. Cardellach, S. Ribó, A. Rius, A. Helm: Detection of arctic ocean tides using interferometric GNSS-R signals, *Geophys. Res. Lett.* **38**(L04103), 4 (2011)
- 40.69 F. Fabra, E. Cardellach, A. Rius, S. Oliveras, O. Nogués-Correig, M. Belmonte-Rivas, M. Semmling, S. D'Addio: Phase altimetry with dual polarization GNSS-R over sea-ice, *IEEE Trans. Geosci. Remote Sens.* **50**(6), 2112–2121 (2011)
- 40.70 E. Cardellach, C.O. Ao, M.G.A. de la Torre-Juárez: Hajj: Carrier phase delay altimetry with GPS-reflection/occultation interferometry from low Earth orbiters, *Geophys. Res. Lett.* **31**(L10402), 1–4 (2004)

- 40.71 A. Helm, H.-U. Wetzel, W. Michajljow, C. Mayer, A. Lambrecht, W. Hagg, A. Dudashvili, G. Beyerle, M. Rothacher: Using reflected GPS signals for the observation of the second 2005 Lake Merzbacher GLOF event, Proc. Int. Workshop Glacial Lake Outburst Floods Central Asia, Bishkek (Central-Asian Institute of Applied Geosciences, Bishkek 2008)
- 40.72 M. Wiehl, B. Legresy, R. Dietrich: Potential of reflected GNSS signals for ice sheet remote sensing, *Prog. Electromagn. Res.* **40**, 177–205 (2003)
- 40.73 E. Cardellach, F. Fabra, A. Rius, S. Pettinato, S. D'Addio: Characterization of dry-snow substructure using GNSS reflected signals, *Remote Sens. Environ.* **124**, 122–134 (2012)
- 40.74 K.M. Larson, J.J. Braun, E.E. Small, V.U. Zavorotny, E.D. Gutmann, A.L. Bilich: GPS multipath and its relation to near-surface soil moisture content, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **3**(1), 91–99 (2010)
- 40.75 N. Rodríguez-Alvarez, A. Camps, M. Vall-Ilossera, X. Bosch-Lluis, A. Monerris, I. Ramos-Perez, E. Valencia, J.F. Marchan-Hernandez, J. Martinez-Fernandez, G. Baroncini-Turricchia, C. Pérez-Gutiérrez, N. Sánchez: Land geophysical parameters retrieval using the interference pattern GNSS-R technique, *IEEE Trans. Geosci. Remote Sens.* **49**(1), 71–84 (2011)
- 40.76 D. Masters, P. Axelrad, S. Katzberg: Initial results of land-reflected GPS bistatic radar measurements in SMEX02, *Remote Sens. Environ.* **92**, 507–520 (2004)
- 40.77 S.J. Katzberg, O. Torres, M.S. Grant, D. Masters: Utilizing calibrated GPS reflected signals to estimate soil reflectivity and dielectric constant: Results from SMEX02, *Remote Sens. Environ.* **100**, 17–28 (2005)
- 40.78 N. Pierdicca, L. Guerriero, R. Giusto, M. Brogioni, A. Egido, N. Floury: GNSS reflections from bare and vegetated soils: Experimental validation of and end-to-end simulator, *Proc. IEEE IGARSS, Vancouver* (2011) pp. 4371–4374
- 40.79 K.M. Larson, E.E. Small, E. Gutmann, A. Bilich, P. Axelrad, J. Braun: Using GPS multipath to measure soil moisture fluctuations: Initial results, *GPS Solutions* **12**, 173–177 (2008)
- 40.80 E.E. Small, K.M. Larson, J.J. Braun: Sensing vegetation growth with reflected GPS signals, *Geophys. Res. Lett.* **37**(L12401), 1–5 (2010)
- 40.81 G. Beyerle, K. Hocke, J. Wickert, T. Schmidt, C. Reigber: GPS radio occultations with CHAMP: A radiographic analysis of GPS signal propagation in the troposphere and surface reflections, *J. Geophys. Res.* **107**(D24), 27.1–27.14 (2002)
- 40.82 G. Foti, C. Gommenginger, P. Jales, M. Unwin, A. Shaw, C. Robertson, J. Roselló: Spaceborne GNSS reflectometry for ocean winds: First results from the UK TechDemoSat-1 mission, *Geophys. Res. Lett.* **42**(13), 5435–5441 (2015)
- 40.83 H. Carreno-Luengo, A. Camps, I. Perez-Ramos, G. Forte, R. Diez R. Onrubia: ³CAT-2: A P(Y) and C/A GNSS-R experimental nano-satellite mission, *Proc. IEEE IGARSS, Melbourne* (2013) pp. 843–846
- 40.84 C.S. Ruf, S. Gleason, Z. Jelenak, S. Katzberg, A. Ridley, R. Rose, J. Scherrer, V. Zavorotny: The CYGNSS nanosatellite constellation hurricane mission, *Proc. IEEE IGARSS, Munich* (2012) pp. 214–216
- 40.85 J. Wickert, G. Beyerle, E. Cardellach, C. Förste, T. Gruber, A. Helm, M.P. Hess, P. Høeg, N. Jakowski, O. Montenbruck, A. Rius, M. Rothacher, C.K. Shum, C. Zuffada: GNSS Reflectometry, Radio Occultation and Scatterometry onboard ISS for long-term monitoring of climate observations using innovative space geodetic techniques on-board the International Space Station. Proposal in response to Call: ESA Research Announcement for ISS Experiments relevant to study of Global Climate Change (GFZ, Potsdam 2011)

41. GNSS Time and Frequency Transfer

Pascale Defraigne

Time and navigation are intimately linked and rely on each other. Global navigation satellite system (GNSS) positioning is based on the measurement of time intervals needed by the signal to travel from satellites to the receiving station on the Earth or nearby. The precision of GNSS positioning is reached thanks to atomic frequency standards onboard the satellites and the possibility to determine their synchronization differences at the subnanosecond level. Time is thereby the core of GNSS. Inversely GNSS is widely used for accurate time and frequency dissemination, as well as for the comparison of distant clocks as needed for time and frequency metrology. All these aspects of using GNSS for time/frequency applications will be presented in this chapter.

41.1	GNSS Time and Frequency Dissemination	1187
41.1.1	Getting UTC from GNSS.....	1188
41.1.2	GNSS Disciplined Oscillators	1189
41.2	Remote Clock Comparisons	1191
41.2.1	The GNSS Time Transfer Technique	1191
41.2.2	Time Transfer Standard CGGTTS	1192
41.2.3	Common View or All-in-View	1193
41.2.4	Precise Point Positioning.....	1194
41.3	Hardware Architecture and Calibration	1197
41.3.1	Time Receivers.....	1197
41.3.2	Hardware Calibration	1198
41.4	Multi-GNSS Time Transfer	1201
41.4.1	General Requirements	1201
41.4.2	GPS + GLONASS Combination	1202
41.4.3	Time Transfer with Galileo and BeiDou	1203
41.5	Conclusions	1203
	References	1204

41.1 GNSS Time and Frequency Dissemination

As already mentioned in Chap. 19, the basis of the global navigation satellite system (GNSS) measurements is the time interval between the emission (satellite) and reception (receiver) of the pseudorandom noise (PRN) codes. The emission time $t_e(\text{sat})$ is read in the satellite clock, while the reception time $t_r(\text{rec})$ is read in the receiver clock. The pseudorange measurement can be denoted as

$$P = c(t_r(\text{rec}) - t_e(\text{sat})) , \quad (41.1)$$

where c is the velocity of light. The satellite clock and the receiver clock being not synchronized, the synchronization error $(t_{\text{rec}} - t_{\text{sat}})$ between those clocks must be taken into account to get the true time interval $(t_r - t_e)$ between the emission and reception, as if it was measured with a same clock

$$(t_r(\text{rec}) - t_e(\text{sat})) = (t_r - t_e) + (t_{\text{rec}} - t_{\text{sat}}) + \text{errors} . \quad (41.2)$$

The true travel time $(t_r - t_e)$ multiplied by the velocity of light corresponds to the true distance between the satellite and the receiver. The pseudorange (41.1) can then be expressed as the sum of a distance, a clock synchronization error, and additional errors mainly due to atmospheric delays, multipath and noise, and hardware delays

$$P = \|\mathbf{x}_s - \mathbf{x}_r\| + c(t_{\text{rec}} - t_{\text{sat}}) + \text{errors} . \quad (41.3)$$

This equation contains four unknowns, three for the position and one additional which is the synchronization error $(t_{\text{rec}} - t_{\text{sat}})$ between the satellite and the receiver clocks. The fundamental of GNSS is to combine observations from several satellites to solve for these four unknowns. However, as the clocks carried by different satellites are not perfectly synchronized, the quantity $(t_{\text{rec}} - t_{\text{sat}})$ is different for all satellites and the total number of unknowns would be $3 + k$, where k is the number of satellites observed at a given epoch. The system

could therefore not be solved. For this reason, the GNSS maintains a reference time scale and provides in the navigation message the quantity $(t_{\text{sat}} - t_{\text{ref}})$ which is the synchronization error of the satellite clocks with respect to this reference. Equation (41.3) hence can be decomposed as follows

$$P = \|x_s - x_r\| + c(\Delta t_{\text{rec}} - \Delta t_{\text{sat}}) + \text{errors} , \quad (41.4)$$

where $\Delta t_{\text{sat}} = (t_{\text{sat}} - t_{\text{ref}})$ is known from the navigation message and $\Delta t_{\text{rec}} = (t_{\text{rec}} - t_{\text{ref}})$ is the synchronization error between the receiver clock and the reference time scale of the GNSS. With this formulation (41.4), Δt_{rec} is the same unknown for all satellites observed at a given epoch, so that the number of unknowns is always four at any time, and the user can determine them continuously provided that a minimum of four satellites are simultaneously visible.

The reference time scale *ref* depends on the satellite clock products used. It is the reference time scale of the constellation when using the broadcast navigation messages, but various time scales are used as references in postprocessed products like those provided by the International GNSS Service (IGS) community (Chaps. 33 and 34).

For all timing applications, the most important information here is $\Delta t_{\text{rec}} = (t_{\text{rec}} - t_{\text{ref}})$, the synchronization error between the receiver and the reference. Getting this quantity for two receivers at a same epoch, whatever the distance between them, provides the difference between the two receiver clocks $(t_{\text{rec},1} - t_{\text{rec},2})$ at that epoch (Sect. 41.2). The present section is dedicated to the time information disseminated by the GNSS, allowing any user to get an accurate time and/or frequency.

41.1.1 Getting UTC from GNSS

Among the various procedures existing for time dissemination, GNSS is certainly the most popular when a sub-millisecond precision is required. It is for example extensively used for precise time tagging [41.1], banking, synchronization of communication and telecommunication networks [41.2], or the phase synchronization energy transport and distribution networks [41.3]. Each base station of the network being synchronized continuously on the accurate time disseminated by the GNSS satellites, this assures the synchronization between all the base stations.

The basis of any official time in the world is the Universal Time Coordinated (Chap. 2); local time and legal time can be directly obtained by adding the time zone corrections to Coordinated Universal Time (UTC). Each of the GNSS provides in its navigation message

a second degree polynomial modeling the evolution of the difference between its reference time scale and a prediction of UTC. Combining this quantity with the Δt_{rec} estimated from the GNSS code measurements gives at each observation epoch the synchronization error between the receiver clock and the prediction of UTC

$$(t_{\text{rec}} - t_{\text{ref}}) + (t_{\text{ref}} - \text{UTC}) = (t_{\text{rec}} - \text{UTC}) . \quad (41.5)$$

This synchronization error can then be applied to the receiver internal clock time to produce a 1 pulse per second (pps) signal synchronized continuously with this prediction of UTC.

It must be noted that the true UTC does not exist in real time. It is indeed computed monthly in postprocessing by the International Bureau of Weight and Measurements (BIPM, Chap. 2). As a consequence, any user requiring precise timing information in real time can only rely on a prediction of UTC. The best predictions of UTC are realized by the time laboratories whose clocks are contributing to the clock ensemble providing UTC. These laboratories maintain a local realization of UTC, named UTC(k) where k is the acronym of the laboratory. After each month, when UTC is computed, the BIPM reports about the differences between each prediction UTC(k) and the true UTC, and their statistical uncertainties; this information is available freely on the BIPM website. This assures the traceability of all the UTC(k) realizations. BIPM recommends that all the UTC(k) realizations be maintained at less than 100 ns of UTC, but a good proportion of laboratories reach the level of some nanoseconds, as illustrated in Fig. 41.1 which shows two examples of UTC realizations, namely UTC (PTB) and UTC (United States Naval Observatory (USNO)), where PTB is the German National Metrology Institute, and USNO is the US Naval Observatory.

Each of the GNSS constellations relies on its own reference time scale and disseminates its own prediction of UTC, which is UTC(USNO) for the Global Positioning System (GPS), UTC(SU) for the Russian Global Navigation Satellite System (GLONASS), an average of five European UTC(k)'s for Galileo and UTC(NTSC) for BeiDou, where NTSC is the National Time Service Center in China. Also Quasi-Zenith Satellite System (QZSS) is providing a link to the prediction UTC(NICT) realized by the National Institute of Information and Communications Technology in Japan. As the link to UTC should be available in real time for the users, it is broadcast as a prediction to be used till the next update. Except for GLONASS, these predictions of UTC are broadcast worldwide with an uncertainty of only a few nanoseconds. This is not the

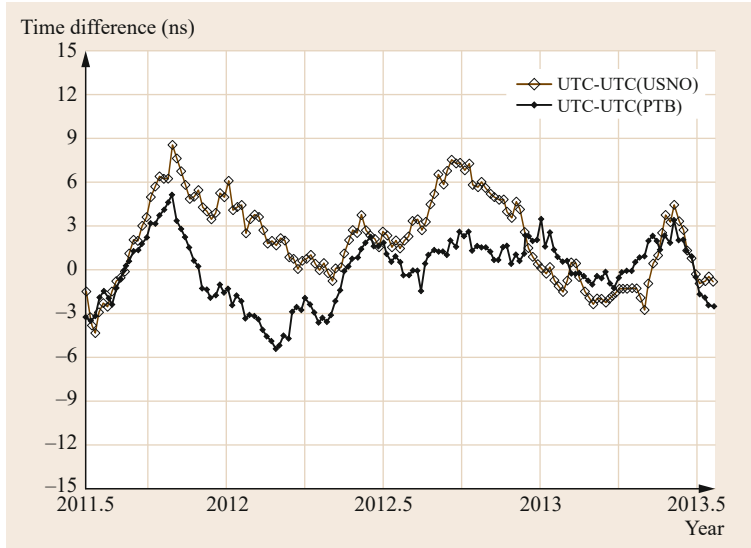


Fig. 41.1 Differences between UTC and the realizations UTC (USNO) and UTC (PTB) over 2 years

case for GLONASS, limited by an uncertainty of hundreds of nanoseconds, but it is likely to be improved in the near future through appropriate calibrations. Since January 2011, the BIPM publishes in Sect. 41.5 of its circular T values of [UTC–UTC(USNO)_GPS] and [UTC–UTC(SU)_GLONASS], that is, the differences between the true UTC and the predictions broadcast by GPS and GLONASS, respectively; during the 2 year period corresponding to Fig. 41.1, the differences [UTC–UTC(USNO)_GPS] remained within [–12, 12] ns, while the differences [UTC–UTC(SU)_GLONASS] varied within [–440, –240] ns.

At the user level, an accuracy of some nanoseconds on the prediction of UTC can thus presently be obtained only with GPS; Galileo and BeiDou will offer the same capability in the near future. However, the nanosecond accuracy can only be reached if the signal delays in the antenna, cable, and receiver are known; these delays reach the level of hundreds of nanoseconds, so that without any calibration of the receiving chain, GNSS can provide an access to UTC with a submicrosecond accuracy. Calibration aspects will be detailed in Sect. 41.3.2.

The accuracy statement here above should be understood within the way the uncertainties are defined for time dissemination and time transfer. As recommended in the guide to the expression of uncertainty in measurement produced by the working group 1 of the Joint Committee for Guides in Metrology [41.4], a distinction is made between the type A and type B uncertainties. The precision is given by the type A uncertainty u_A , which corresponds to the statistical uncertainty, evaluated by taking into account the level of phase noise in the raw data and the magnitude of effects

varying over a typical duration below one month [41.5]; the type B uncertainty u_B is the uncertainty of the calibration. The uncertainties u_A and u_B correspond to the variance and bias in a classical decomposition of the mean-square-error. The accuracy of time measurements is then given by the combined uncertainty

$$u = \sqrt{u_A^2 + u_B^2}.$$

41.1.2 GNSS Disciplined Oscillators

For all applications requiring precise and stable frequencies, the expensive purchase of an atomic clock can be substituted by the use of a GNSS-disciplined oscillator (GNSSDO in what follows), for which the cost is much less than cesium standards.

As explained in the previous section, GNSS signals allow one to determine continuously the synchronization error between the local receiver clock and the prediction of UTC, with an accuracy of 1 μ s or better, and a precision of a tens of nanoseconds. The GNSS receiver can therefore output a 1 pps signal synchronized continuously with one of the best available realizations of UTC. Considering a precision of 10 ns on the timing, it is theoretically possible to reach a frequency stability of $1 \cdot 10^{-13}$ at an averaging time of one day.

The principle of GNSSDOs (Fig. 41.2) is to generate time and frequency signals using a voltage-controlled oscillator (VCO), which can be a high-quality quartz or a rubidium oscillator, and whose frequency is controlled by timing information broadcast by the GNSS satellites and reproduced by the

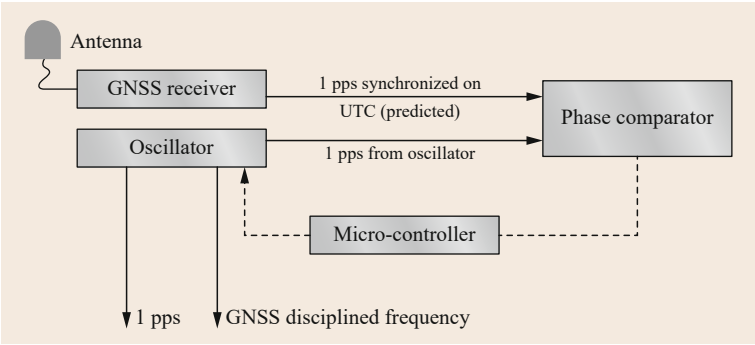


Fig. 41.2 Schematic representation of the GNSS-disciplined oscillator

GNSS receiver in its 1 pps output. The local oscillator is controlled with a servo loop, in a similar way as a phase-locked loop (PLL, see Chap. 13). In its basic form, the PLL compares the phase of the reference signal given by the GNSS receiver to the phase of the oscillator. The phase detector then outputs the phase difference between the two input signals, and a microcontroller sends the correction to be applied to the oscillator to be aligned with the GNSS received signal. In some cases, the software used by the microcontroller compensates for not only the phase and frequency changes of the local oscillator, but also for the effects of aging, temperature, and other environmental parameters [41.6]. These effects being modeled can still be corrected for in the case of temporary interruption of GNSS signals.

The software also provides the ability to vary its time constant. For example, if a more stable oscillator is used, the software can adapt the servo loop to use a longer time constant and make frequency corrections less often. Indeed, thanks to their permanent synchronization with UTC, GNSS receivers have excellent long-term stability at averaging times greater than several hours. However, their short-term stability is degraded by the noise in GNSS signals, due to multipath, atmospheric perturbations, and uncertainties in orbit and clocks broadcast in the navigation message, conferring a precision of about 30 ns on the 1 pps. On the other hand, a rubidium or a good quality quartz oscillator (like an oven-controlled oscillator) has better short term stability but is susceptible to long-term effects like aging. A GNSSDO aims at using the best of both sources, combining the short-term stability performance of the oscillator with the long-term stability of the GNSS signals to give a reference source with excellent overall stability characteristics. The time constant of the steering is therefore chosen as a function of the stability of the oscillator, compared to the stability of the GNSS-based frequency.

An illustration is provided in Fig. 41.3 which presents the Allan deviation (giving the frequency sta-

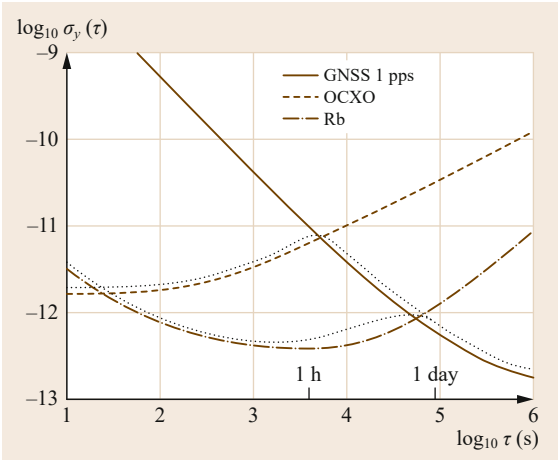


Fig. 41.3 Frequency stability of the GNSSDO for two distinct types of oscillators. The solid line corresponds to the Allan deviation of the GNSS, dashed lines to the Allan deviation of the oscillators, and dotted lines correspond to the Allan deviation of the frequency delivered by the GNSSDO. The GNSS 1 pps frequency stability is based on a 1 pps precision of about 30 ns

bility of the signal on different averaging times, as defined in Chap. 5) of two distinct oscillators in comparison with the Allan deviation of the GNSS-based timing signals (solid line). A time constant of about 1 h would be chosen for the illustrated oven-controlled crystal oscillator (OCXO) as for longer averaging times the GNSS-based frequency has a better stability. For the Rubidium oscillator, a time constant of about one day would be preferred to keep the frequency stability of the oscillator for shorter averaging times as it is there better than the stability of the GNSS-based frequency. For long averaging times, the GNSSDO is always based on the GNSS-based frequencies, whatever the oscillator used.

The performances of GNSSDOs are highly variable among the available models, as influenced by the design characteristics, by the oscillator implemented, by

the number of frequencies measured by the receiver (dual-frequency receiver allowing for ionospheric delays correction), etc. However, any GNSSDO that is locked to the satellite signals should be able to provide, when averaging for periods of several days or longer, a frequency accuracy at the level of some parts in 10^{13} . GNSSDO indeed rely on the GNSS predictions of UTC, and hence provide an accurate frequency (i. e., in agreement with the SI second realized by the UTC), and a long-term stability better than any free running oscillator, including the atomic standards.

GNSSDO are nowadays widely used as primary standards in calibration laboratories. They can theoretically assure the traceability of generated time and frequency to the SI second realized by UTC. Here,

traceability of a measurement to the SI unit is defined as maintaining an unbroken chain of calibrations that trace back to the SI unit. GNSSDO serve as self-calibrating standards that should not require adjustment or calibration. The uncertainties on the time and frequencies generated by the GNSSDO can be determined using firstly the uncertainty of the GNSS measurements, and secondly the differences between the predicted UTC and the true UTC, as published in the BIPM Circular T. Additionally, all signal delays in internal circuits and antenna as well as their uncertainties should be determined if the GNSSDO is used as a time reference. However, as the rules for traceability vary from country to country, it is recommended to the users to refer to their national metrology institute.

41.2 Remote Clock Comparisons

People are familiar with the saying *if you have one clock you always know what time it is, but if you have more than one, you never know!* But in reality, if you have only one clock, you cannot know anything about its accuracy unless you compare it with a second clock whose you know the accuracy, and similarly for their frequency stability; furthermore if your clock suddenly stops to run, you lose completely any time information. This is the reason why it is always recommended to work with several clocks, to monitor continuously the differences between their readings and an accurate time available somewhere. Atomic clock comparisons are also essential for the needs of time and frequency metrology. First, for the generation of UTC, the only data that can be used for clocks are the differences between two clocks at successive epochs. The ensemble algorithm used to produce UTC, or any other time scale, treats therefore the differences between all the clocks of the ensemble. Second, in order to determine the frequency accuracy and the frequency stability of commercial or experimental clocks, these must be compared with other clocks whose frequency accuracy and/or stability is at least the same as the one of the clock examined. Finally, clock comparisons can be necessary for scientific applications requiring a measurement of a time interval of which the starting and closing points are measured with different clocks. This is, for example, the case for the measurement of the velocity of the neutrinos [41.7] where the clock measuring the departure time and the clock measuring the arrival time are separated by some hundreds of kilometers.

In a local environment, clocks can be compared using a phase or frequency comparator, or a time interval counter, while for remote clocks separated by some thousands of kilometers other techniques must be en-

visaged. A prime requisite is that the methods of comparison at a distance do not contaminate the frequency stability of the clocks. For applications requiring only a moderate precision (some hundredths of a second), any clock can be compared to the precise time delivered by some precise time facility via the internet and the network time protocol, an internet-based hierarchical time transfer technique [41.8], or to the timing signals delivered by some radio-frequency emitting station connected to a UTC(k) realization. However, these precisions are often insufficient and the use of satellite systems is necessary; one of them is the GNSS.

41.2.1 The GNSS Time Transfer Technique

The technique, called GNSS time transfer, was used since the 1980s for the comparisons of remote frequency standards needed for the realization of UTC [41.9]. Additionally, as a consequence of its reduced cost, the technique is widely used by private companies offering time stamping, or time and frequency calibration. They then continuously monitor their local clock against the realization of UTC maintained by their national metrology institute.

GNSS time transfer is based on the following principle illustrated in Fig. 41.4. The first step is to determine the synchronization error between the local clock and the reference time scale of the GNSS. This is achieved by connecting each clock to a GNSS receiver, in such a way that the synchronization error between the internal receiver clock and the external clock can be continuously measured. From the GNSS signals, each receiver determines

$$\Delta t_{\text{rec},i} = (t_{\text{rec},i} - t_{\text{ref}})$$

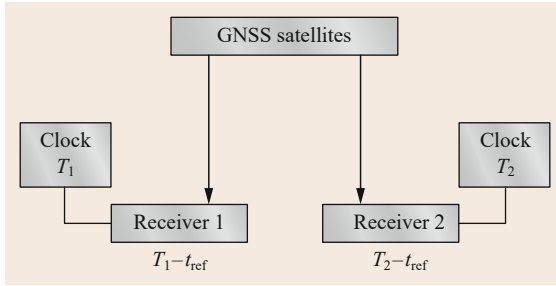


Fig. 41.4 Schematic principle of GNSS time transfer, that is, remote clock comparison

as explained in Sect. 41.1, and from the external measurement between the receiver and the clock T one gets in each laboratory

$$(T_i - t_{ref}) = (t_{rec,i} - t_{ref}) - (t_{rec,i} - T_i). \quad (41.6)$$

The second step consists in computing the difference between the quantity $(T_i - t_{ref})$ obtained in two stations ($i = 1, 2$) from simultaneous observations to get $(T_1 - T_2)$, that is, the synchronization error between the two remote clocks. Note that in the present case simultaneous means that the two observation epochs should not differ by more than one microsecond, which can be easily reached with any GNSS receiver.

In what follows, different analysis strategies as well as the instrumental setup and requirements needed to enable a stable and accurate time and frequency transfer are discussed.

41.2.2 Time Transfer Standard CGGTTS

As timing information can only be provided by the GNSS code measurements, due to the ambiguities inherent to the carrier-phase data, the most employed time transfer tools are based on code measurements only. Carrier-phase data however provide high precision frequency comparisons, and the use of precise point positioning (Chap. 25) for time transfer is currently widespread (Sect. 41.2.4), and used for the computation of UTC since 2009 [41.10].

The common GNSS generic time transfer standard (CGGTTS) has been developed by the Consultative Committee of Time and Frequency (CCTF) as a common format to facilitate the data exchange for time dissemination and time transfer. The last version V2E [41.11] covers the use of GPS, GLONASS, Galileo, BeiDou, and QZSS and has evolved from an earlier GPS-only standard. CGGTTS files contain, among other associated quantities, the differences between the clock connected to the GNSS receiver and the GNSS reference time scale $(T - t_{ref})$. These differences

result from a well-defined analysis procedure of code measurements [41.12] in which the station coordinates are fixed and the satellite position and satellite clock are extracted from the broadcast navigation messages. The computation procedure applies to satellite tracks of 13 min. For each satellite visible during this 13 min period, the corresponding solution $(T - t_{ref})$ is reported in the CGGTTS file. The tracking schedule is distributed by the BIPM as a list of the starting epochs of the tracks. Note that this 13 min duration was decided in the 1980s as it was the time required by a receiver to acquire a full GPS navigation message.

The common-view method was proposed in the 1980s by Allan and Weiss [41.9] and the associated CCTF format was based on one-channel C/A code receivers. Following the improvements of atomic frequency standards in terms of precision and accuracy, GPS (or more generally GNSS) time and frequency transfer underwent major evolutions both at the algorithmic levels and at the hardware level. A first improvement was found in the use of a multichannel approach [41.13], increasing the number of satellites which reduces correspondingly the noise of clock solutions. For applications requiring the highest precision, as for example the computation of international atomic time (TAI, Chap. 2), the CGGTTS results are improved by adding a correction for satellite orbits and clocks using the rapid IGS products. Also, the ionospheric correction used in the CGGTTS results, based on the broadcast ionospheric model of the constellation, is replaced by a new estimation based on IONosphere map EXchange format (IONEX) maps (Annex) delivered by the IGS [41.14]. A further upgrade of the CGGTTS was the use of dual-frequency receivers measuring the GPS P(Y)-codes, enabling to remove the ionosphere delays at the first order, and leading to a factor-of-2 improvement in the precision of the intercontinental time links [41.15]. Note that for short baselines, the increase of noise in the ionosphere-free combination with respect to the single-frequency time transfer solution can be larger than the residual ionospheric errors associated with the Klobuchar model or with the IONEX maps. The same ionospheric delay is indeed suffered by the GNSS signals when they arrive in stations close to each other. However, the timing community is preferring using the ionosphere-free combination so that the CGGTTS files can be used easily whatever the distance of the second clock entering into the comparison may be.

An example of CGGTTS file is shown in Fig. 41.5. The header of the file summarizes the station information, that is, receiver name, station coordinates, and hardware delays (Sect. 41.3.2) used for the com-


```

CGGTTS      GENERIC DATA FORMAT VERSION = 2E
REV DATE = 2015-02-20
RCVR = RRRRRRRR
CH = 12
TMS = IIIIIIII
LAB = ABC
X = +4027889.79 m
Y = +306995.67 m
Z = +4919491.36 m
FRAME = ITRF
COMMENTS = NO COMMENTS
INT DLY = 53.9 ns (GPS P1), 49.8 ns (GPS P2)    CAL_ID = 1nnn-yyyy
CAB DLY = 200.0 ns
REF DLY = 120.6 ns
REF = UTC(ABC)
CKSUM = 3B

```

SAT	CL	MJD	STTIME	TRKL	ELV	AZTH	REFSV	SRSV	REFSYS	SRSYS	DSG	IOE	MDTR	SMDT	MDIO	SMDI	MSIO	SMSI	ISG	FR	HC	FRC	CK
			hhmmss	s	.ldg	.ldg	.lms	.lps/s	.lms	.lps/s	.lms		.lms	.lps/s	.lms	.lps/s	.lms	.lps/s	.lms				
G24	FF	57000	000600	780	317	394	+1186342	+0	163	+0	40	12	141	+22	23	-1	23	-1	29	+2	0	L3P	5C
G05	FF	57000	000600	780	70	2325	+22617	+6	165	-3	53	26	646	+606	131	-9	131	-9	37	+1	0	L3P	8C
G17	FF	57000	000600	780	509	1217	-1407831	-36	154	-54	20	31	100	-8	24	+0	24	0	13	+4	0	L3P	7A
G16	FF	57000	000600	780	300	3022	+308130	-18	246	-28	29	41	134	-22	63	+4	63	4	21	-1	0	L3P	80

Fig. 41.5 Example of CGGTTS file

putation. The results are then provided, with each line corresponding to one satellite 13 min track. The columns for the time transfer solutions are REFSV and REFSYS, that is, the differences modulo one second between the laboratory clock and the satellite in view (SV) or the system time scale (SYS). They correspond to the midpoint of a linear fit applied to the 13 min results; the standard deviations with respect to this linear term are also provided (columns SRSV and SRSYS).

41.2.3 Common View or All-in-View

The initial CGGTTS files were produced by single channel receivers and the time transfer was computed as the differences of the CGGTTS results collected simultaneously from the same satellite by the two stations. The technique received the name of GPS common view (as at that time only GPS was used). All the satellite hardware delays or satellite clock errors are removed by this technique; the remaining errors are mainly due to different atmospheric distributions on the signals received at the remote stations, and the multipath at the stations. This common view (CV in what follows) technique was also used when multichannel receivers entered into the time laboratories. The final time transfer solution for the clocks T_1 and T_2 is then for each 13 min track of the BIPM schedule, a weighted average of the results obtained with the satellites in common view of both stations

$$(T_1 - T_2)(t) = \frac{1}{N(t)} \sum_{i=1}^{N(t)} w_i [(T_1 - t_{\text{ref}})_i(t) - (T_2 - t_{\text{ref}})_i(t)], \quad (41.7)$$

where $(T_x - t_{\text{ref}})_i(t)$ is the solution found in the CGGTTS file from station x for the satellite i at the epoch t , w_i is the weight, generally the $\sin^2(E)$ with E the satellite elevation, and $N(t)$ is the number of satellites simultaneously visible by the two stations. However, the quality of the CV solutions tends to degrade with increasing distance between the stations, since the number of simultaneously observed satellites decreases as the baseline increases.

An alternative to the common-view technique is therefore called the *all-in-view* (AV) approach: a clock solution $(T_x - t_{\text{ref}})(t)$ is computed independently for each station using all visible satellites and the difference is then computed afterward

$$(T_1 - T_2)(t) = \frac{1}{L(t)} \sum_{i=1}^{L(t)} w_{i1}(t)(T_1 - t_{\text{ref}})_i(t) - \frac{1}{M(t)} \sum_{i=1}^{M(t)} w_{i2}(t)(T_2 - t_{\text{ref}})_i(t), \quad (41.8)$$

where $L(t)$ and $M(t)$ are the total number of observed satellites by stations 1 and 2, respectively, at the epoch t as in (41.7). AV is therefore independent of the distance between the stations. This is the same principle as precise point positioning (Sect. 41.2.4) but using only code measurements and fixing the position to its known value. Of course, the errors from satellite clock or ephemeris estimate do not cancel as they do in the common-view technique. Therefore, the use of precise ephemerides and clocks rather than the broadcast navi-

gation messages is mandatory [41.16]. Using IGS rapid products, the remaining uncertainties due to satellite orbits and clocks average appropriately to well below 100 ps for averaging 1 d and longer [41.17]. These authors also demonstrate the superiority of AV with respect to CV for baselines longer than 2000 km.

The choice between the CV and the AV will therefore rely on the distance between the stations where the clocks are located, and also the availability of precise orbits and clocks. CV will be preferred when only the broadcast ephemerides are available, but this approach should be restricted to short distance clock comparisons.

As the time transfer based on CGGTTS is a code-only analysis, both AV and CV are significantly affected by multipath of the code signals as well as uncertainties of the hardware delays. Depending on the station setup, some important diurnal variations can appear in the time transfer solution, which are not a clock variation but only the signature of the code multipath in one or both stations. An example is provided in Fig. 41.6, where specific patterns appear in the clock solution with a 23 h 56 min periodicity, that is, the needed time to retrieve the same geometrical relationship between the satellite, the receiving antenna, and the nearby reflectors.

To date, all in view time transfer using the ionosphere-free combination of GPS L1/L2 P(Y)-code observations constitutes the state of the art in GNSS time transfer using code measurements only. The statistical uncertainty (u_A) is at the level of a few nanoseconds, being limited by the current noise and multipath of the code measurements. The systematic uncertainty (u_B) relies on the calibration capabilities described in Sect. 41.3.2.

41.2.4 Precise Point Positioning

The noise and multipath of the code measurements in a code-only analysis masks the short-term stability of some atomic clocks, for example, hydrogen masers. A significantly higher stability can be obtained by us-

ing the carrier-phase measurements in addition to the code data. This requires a combined analysis of both code and carrier-phase measurements with a consistent modeling of these measurements similar to GNSS data analysis dedicated to precise positioning.

Only processing zero differences (e.g., precise point positioning) or single differences can be used for time transfer because the receiver clock disappears in double differences. Single differences rely on the same principle as the code-only CV approach, but using both code and carrier-phase measurements, while precise point positioning (PPP) relies on the same principle as the AV. As in the choice between CV and AV, PPP is usually preferred to single-difference analysis as independent on the baseline length. The impact of the satellite geometry on the single-difference solution was clearly demonstrated, for example, in [41.18], where differences between single-difference and PPP analyses reach the nanosecond level for an intercontinental baseline.

As explained in Chap. 25, PPP provides, besides the station position (static or kinematic), the tropospheric delay and the receiver clock solution. When used for time and frequency transfer, the station is of course considered as static, while the receiver clock is solved for each observation epoch. The receiver clock solution is, as explained before, $(t_{\text{rec}} - t_{\text{ref}})$ where t_{ref} is the reference time scale of the satellite clock products used in the PPP processing. The time transfer solution, as explained in Sect. 41.2, is then a difference between the clock solutions obtained for two remote stations.

It is mandatory that both PPP clock solutions have been computed using the same satellite orbit and clock products so that the reference is the same for both. Note that the IGS also provides receiver clock solutions $(t_{\text{rec}} - t_{\text{ref}})$ for part of the stations in the network. These are computed using zero differences, but the satellite and station clocks are computed at the same time, fixing the tropospheric delays to the values determined from a double difference network processing. As the IGS solution is based on the combination of solutions obtained by different analysis centers, it is usually considered as

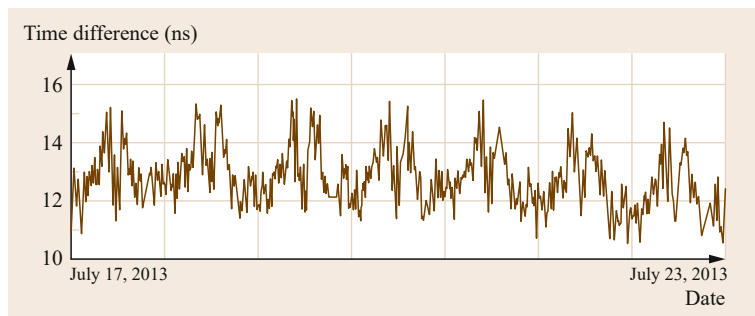


Fig. 41.6 Example of multipath impact in GNSS time transfer based on CGGTTS: clock comparison between the time laboratory ROA (Spain) and PTB (Germany), computed with the all in view technique based on the ionosphere-free combination of GPS P(Y)-code measurements on the L1 and L2 frequencies

the best solution available for the stations included in the network. The reference time scale of these solutions is the IGS time scale IGST (resp. IGRT) for the final (resp. rapid) products.

In a time transfer solution, the shape of the curve corresponds to the frequency variations between the two clocks (or timescales) compared, while the position of the curve on the y-axis corresponds to the time synchronization difference between the two clocks (or timescales). When the solution is computed using a combined analysis of code and carrier-phase measurements, the shape of the curve will be given by the carrier-phase data, while its position on the y-axis will be given by the code measurements. The carrier-phase data indeed contain an ambiguity term, which would put the solution on an arbitrary value on the y-axis. This ambiguity is determined for each satellite continuous track from the differences between the code and the carrier-phase measurements, and the final solution is then given by the carrier-phase data corrected for their ambiguity term. Thanks to their higher precision, carrier-phase data improve significantly the comparison of remote clock frequencies with respect to a code-only solution.

This improvement is illustrated in Fig. 41.7 for the comparison of two masers located in Brussels (Royal Observatory of Belgium) and Braunschweig (Physikalisch-Technische Bundesanstalt); both AV and PPP solutions are presented as well as their frequency

stabilities highlighted by the Allan deviation. The statistical uncertainty u_A of PPP time transfer is currently below 100 ps for each observation epoch, allowing frequency transfer with an uncertainty approaching $1 \cdot 10^{-15}$ or even better for averaging times of one day [41.19–21]. The systematic uncertainty u_B of PPP time transfer is however the same as for code-only solutions, that is, a few nanoseconds, and relies on the calibration capabilities detailed in Sect. 41.3.2.

In a PPP analysis, the noise of the code measurements is responsible for jumps between successive and independent clock solutions. Indeed, as explained here above, the carrier-phase ambiguities are determined as the average over the continuous satellite track of the differences between the carrier-phase and code pseudoranges (maybe corrected with some bias to assure the integer nature of the ambiguity). As a consequence, the absolute values of the final PPP solution, that is, position of the curve on the y-axis, correspond roughly to the average of the code measurements of the data batch analyzed. Due to the noise of the code measurements, the standard error of the mean (SEM) is not zero and one can have jumps between two successive clock solutions. Consider, for example, one day of data sampled at 5 min with four visible satellites at each epoch. For a pseudorange pure white noise with an standard deviation of 30 cm, the standard error of the mean clock solution is then typically about 33 ps. This implies jumps between the successive daily clock solutions, distributed as a white noise with a zero mean and a standard deviation of 47 ps. However, the magnitude of these day-boundary jumps can be significantly larger [41.22]. The standard deviations of the jumps are station dependent, ranging from 150–1000 ps [41.23], well larger than the expected 47 ps. Only stations equipped with H-masers are considered there as for less stable frequency standards the clock instability dominates the day-boundary jumps caused by the pseudorange noise.

The origin of these large day-boundary jumps and of their station-dependent behavior is not yet fully understood, but reflects different station code performances, and reveals the colored signature of the code measurements. Several causes have already been identified, for example, a correlation with external temperature variations, but with opposite sign for different stations [41.22]. A large part of the day-boundary jumps (especially those of large magnitude) was also shown to be associated with pseudorange variations similar to all the satellites [41.24], with magnitudes reaching several nanoseconds, and possibly caused by instrumental delay variations (which can be due to temperature variations), reflections in the cable connectors or some nongeometrical near-field effect.

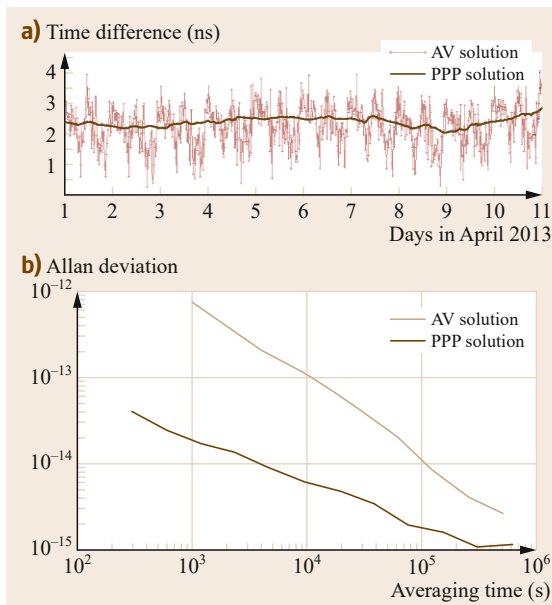


Fig. 41.7a,b Comparison of the time transfer solutions between two H-masers located in Brussels, Belgium and Braunschweig, Germany, computed with either AV or PPP (a) and associated Allan deviations (b)

As an example, Fig. 41.8 presents the PPP clock solution for the IGS station OPMT (Paris) equipped with a H-maser. Based on the use of IGS final products, the solution corresponds to (OPMT-IGST). A same linear term was removed from all the curves in order to facilitate the visibility. The results depicted in Fig. 41.8 correspond to (1) a PPP solution from daily processing, (2) a PPP solution from a unique process of a one month data batch (computed by the BIPM for the computation of UTC), and (gray dots) a code-only solution plotted for each satellite separately. The observed day boundary jumps in curve (1) are because the absolute value of the PPP solution corresponds to the average of the code measurements of the data batch analyzed, while these code measurements suffer from some long-term variations, as seen from the code-only solution (gray dots). For comparison, the PPP solution computed for the complete five day data batch is of course continuous across the day boundaries. Note that the two PPP solutions presented in Fig. 41.8 have been produced by two different software tools, which explains the small differences in the subdiurnal variations.

Several approaches have been addressed in order to reduce or eliminate the daily discontinuities. Correcting the jumps observed in the clock solution when the ambiguities are still float in the solution produces a random walk of the continuous solution [41.20]. Processing multiday data batches [41.25] as proposed in curve (2) of Fig. 41.8, transports of course the problem at the batch boundaries, but in that case the jumps are smaller as the impact of the code noise is reduced thanks to the increased number of observations. A ded-

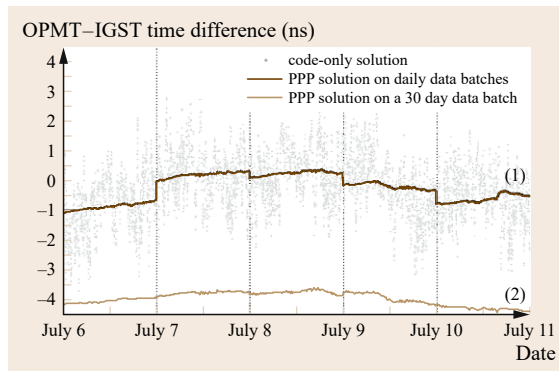


Fig. 41.8 PPP solutions for the station OPMT (Paris) equipped with a H-maser. Solution (1) is based on daily process, while solution (2) was obtained from the analysis of a one month data batch. Gray dots are the code-only solutions plotted for each satellite separately. In order to improve the visibility, a same linear term was removed from all the curves as well as a 3 ns bias from curve (2)

icated data filtering method [41.26] was also proposed, as well as processing sliding windows [41.27] but again the problem is mitigated rather than solved.

The optimal way to produce independent solutions while ensuring their continuity is to fix the ambiguities to integer values [41.28]. This requires the introduction of some station and satellite biases to absorb the non integer part of the ambiguities. In the results produced by this integer PPP processing, the day boundary discontinuities still exist but are always an integer number of cycles of the narrow-lane combination of the two frequencies used in the ionosphere-free combination; these integer jumps can then be easily canceled out [41.29]. Equivalently, in [41.30] a new parametrization is proposed that separates the pseudorange observation colored noise from the carrier-phase parameters, that is, ambiguities and clock; the continuity is then directly obtained between independent PPP solutions of successive data batches, but requires to arbitrarily fix one initial receiver bias.

All these issues related to the continuity of the PPP solutions at the batch boundaries and to the reduction of the impact of the noisy code measurements find their importance in the wish to improve the performances of GNSS frequency transfer. The time transfer quality however always relies on the code measurements, and the best way to improve these measurements will be to design an antenna setup which reduces the near-field multipath. Using either absorbing material around the antenna or a pillar of at least 2 m height supporting the antenna, have for example, shown convincing results in reducing the magnitude of the day boundary jumps [41.31], or [41.32]. Furthermore, antenna cables with low sensitivity to temperature variations, and a stabilized temperature around the clock and receiver are as well recommended to reduce pseudorange coloured noise.

A last point to be emphasized is the correlation between the clock and the tropospheric zenith path delay (TZD) estimated in the PPP processing. It was, for example, demonstrated that a short sampling rate (shorter or equal to 15 min) in the estimation of the TZD is mandatory to reproduce as much as possible the true troposphere variations and hence to avoid any contamination of the clock solution from unmodeled short-term tropospheric changes [41.33]. The details of the mapping function employed have however no significant impact [41.34].

The ultimate performances of GNSS frequency transfer can be estimated using a pair of stations in a common-clock setup, that is, both connected to a same clock, so that the solution is not influenced by the clock instability. The Allan deviation of such solutions

are presented in Fig. 41.9. The common-clock results were obtained using two separate receiving chains with a distance of 100 m between the antennas. In order to emphasize the impact of the code measurements on the stability of the solutions, Fig. 41.9 presents, for this common-clock setup, one classical PPP solution and one solution obtained using only carrier-phase data. In latter, the single differences of carrier-phase data were used and the ambiguities were determined with respect to zero (the expected clock difference) rather than with respect to the code single differences as is the case in PPP. Therefore, only the noise of the carrier-phase data influences the solution. Note that for nearby stations, processing single differences or PPP provides similar results as PPP because of the geometry of satellites which is exactly the same for both stations. The impact of any error on satellite products has therefore exactly the same impact on the PPP solutions of both stations, and cancels out in the difference of the PPP solutions while in single differences it cancels out at the level of observations. The second curve results from a PPP analysis using the NRCAN PPP software on a multi-day basis [41.25] in order to avoid a contamination of the Allan deviation from the day-to-day discontinuities inherent to daily processing of PPP. The difference between these two curves comes by using code measurements in the PPP case, which degrades the stability at intervals of a few hours, that is, the classical duration of the satellite visibility on which the ambiguities are constant.

The two other curves of Fig. 41.9 present the Allan deviation of the PPP solutions for the links Brussels–Washington (BRUS-USN3, about 6000 km) and Brussels–Paris (BRUX-OPMT, about 300 km). Both provide approximately the same quality. The short-term stability is however lower than what was expected from the common-clock results; the origin of

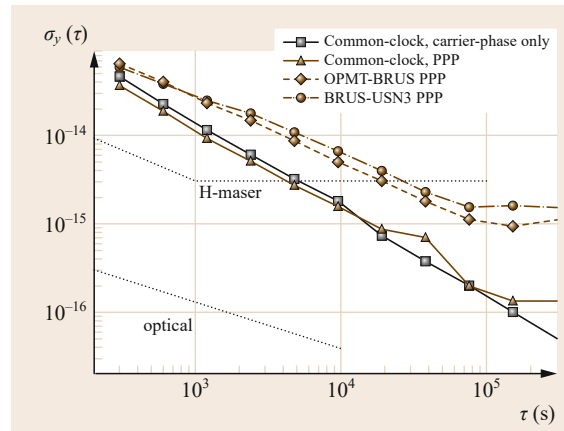


Fig. 41.9 Allan deviation of the clock solutions estimated with state-of-the-art GNSS frequency transfer compared with the Allan deviation of the most stable atomic clocks to date, that is, the H-masers and optical frequency standards. The clock solutions correspond to two links among stations equipped with H-masers: Brussels–Paris (BRUS-OPMT, 300 km), Brussels–Washington (BRUS-USN3, 6000 km); the 100 m baseline is using the same clock for both stations so that the solution does not depend on the stability of the clock and shows the maximum capabilities of the method

this degraded quality has not yet been identified to date. The H-maser stability curve in the figure shows that H-maser instabilities dominate over periods longer than 3 h, so that the curves (3) and (4) associated with H-maser comparisons cannot provide information about the performance of the technique; only the common-clock setup can be used to this end. The optical clock stability curves show that optical clock comparisons from GNSS will be possible only for averaging times longer than several days.

41.3 Hardware Architecture and Calibration

The GNSS equipment needed for time and frequency transfer consists of a receiving antenna connected via cable to a dedicated GNSS receiver and some cable link between the receiver and the external clock to be examined. For time transfer, it is also necessary to have access to the 1 pps given by the external clock, and each of the just mentioned components should be calibrated, which means that the hardware delay of each signal in these instruments or cables should be accurately determined. This section describes the characteristics of dedicated GNSS receivers for time/frequency transfer,

and the existing methods for calibrating the receiving chain.

41.3.1 Time Receivers

Specific GNSS receivers have been developed and commercialized for time transfer. The receiver system consists of an input for an external frequency reference (typically 5 or 10 MHz) to be used in all internal oscillator functions, an input for an external pulsed signal related to the external clock 1 pps, and possibly an in-

ternal time-interval counter. The components may be integrated into a single package or may be separate and connected together by appropriate cables.

Figure 41.10 shows three types of receivers satisfying these requirements. The epoch of the receiver clock can be either:

- Based on the GNSS signals themselves and continuously monitored against the 1 pps signal of the external clock using a time-interval counter (R1 and R2).
- Locked directly to the 1 pps signal from the external clock (R3).

The classical geodetic receiver (R1) can be used for time transfer only if the 1 pps output is related to the internal reference (or receiver clock), and if the relation between the internal reference and the 1 pps output is perfectly known: it must be provided by the manufacturer or measured by the user following a given procedure which is different for each receiver make. The time interval counter (TIC) for R1 and R2 measures the synchronization error between the receiver internal clock and the external clock to be examined. This TIC should measure time intervals (of up to 1 s if required) with a u_B uncertainty approaching 100 ps or better, and a noise level below 100 ps; the TIC measurements should furthermore be reported separately from the GNSS measurements. The main difference between the receiver types R1 and R2 is that the TIC and the computation of the CGGTTS data are inside the receivers R2 while they are external to the classical geodetic receivers R1; for these receivers, the CGGTTS files will be generated using an external software tool that com-

bines the raw measurements available in, for example, Receiver INdependent EXchange format (RINEX (Annex A.1.2) files and the TIC measurements.

In order to overcome the possible noise introduced by the TIC, some geodetic + time receivers (R3) directly synchronize their internal clock (modulo one constant bias) on the external clock to be compared. The user must in that case ensure that the 1 pps signal is coherent with the frequency reference and maintained sufficiently close to the GNSS time scale to assure proper operation. The input 1 pps allows the receiver to choose without ambiguity one particular cycle of the input frequency to form its internal time reference. The receiver clock is so locked in phase on some given point of the input frequency following the pulse of the input 1 pps signal. This kind of receiver cancels the need for a time interval counter and hence provides a final clock solution which is less noisy than the solutions obtained with R1 or R2. Furthermore, the CGGTTS results can be obtained directly from raw measurements available in the RINEX files, using a dedicated software tool as proposed, for example, in [41.35]. In receiver type R3, the internal reference is obtained either by locking the internal oscillator on the external frequency, or by using directly the external frequency for the internal reference. If the internal oscillator is locked on the external frequency with an enslavement system, then the system must be described in full details by the manufacturer to allow for accurate calibration (Sect. 41.3.2). This system must furthermore be designed to introduce no noise on the frequency; adding noise would make impossible the study of frequency stabilities of the best frequency standards via GNSS frequency transfer. Furthermore, the way the internal reference clock is obtained from the external 1 pps must also be described by the manufacturer; this is mandatory to have access to the delay between the external clock and the GNSS measurements, and hence to correctly transfer time.

The Consultative Committee for Time and Frequency (CCTF) advocated in its recommendation S5(2001) [41.36] that the manufacturers of receivers used for timing with GNSS implement the technical guidelines for receiver hardware compiled by the CCTF group on GNSS time transfer standards (CGGTTS). These guidelines have been compiled with the aim of achieving a system that can transfer time with an accuracy of 1 ns or better. A detailed review and extension to new systems can be found in [41.37].

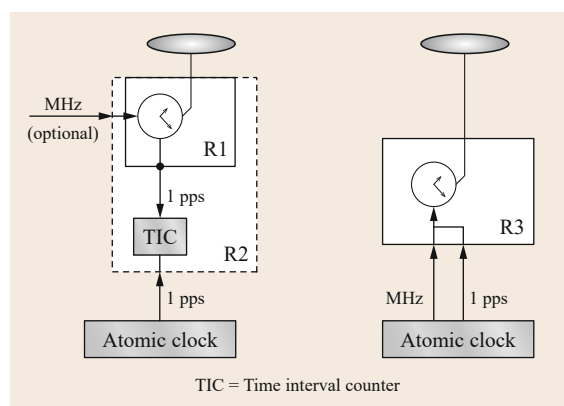


Fig. 41.10 Different kinds of receiver setups for GNSS time and frequency transfer. R1 and R2 use their own internal clock and compares it with the external clock using a time interval counter, while R3 directly uses the 1 pps of the external clock as internal reference

41.3.2 Hardware Calibration

As already stated, GNSS measurements can be exploited for time transfer only if the electric delay accumulated by the signal between the antenna phase

center and the internal timing reference of the receiver is accurately known, as well as the synchronization error between this internal timing reference and the external clock to be examined. Note that the satellite hardware delay being the same for all the ground observing stations, it is already included in the satellite clock and need not be corrected for in the GNSS analysis dedicated to time transfer. One exception comes however when the code measured by the receiver are not the same as the codes used for the satellite clock determination. This happens, for example, when the receiver measures the GPS L1 C/A code. All the clock products for GPS satellites are indeed based on the ionosphere-free combination of L1/L2 P(Y)-codes. Using the combination of L1 C/A and L2 P(Y) code measurements requires therefore a transformation of the satellite clock products to the same combination, using the satellite differential code biases L1 C/A–L2 P(Y) that are provided, for example, by the IGS (Sects. 19.6.1 and 21.3.1).

The station hardware delays, in contrary, must be determined by calibration, as well as the exact time offset between the receiver internal clock and the external clock. All these delays are represented schematically in Fig. 41.11 for the three types of receivers described earlier.

Hardware delays exist in both code and carrier-phase measurements. However, only the code delays are determined by calibration and corrected for in time transfer computation, since only the code measurements provide the time. When carrier-phase data are used, the phase delays are absorbed in the ambiguities.

The first category of delays consists of the electric delays affecting the GNSS signals, that is, δ_A the antenna delay, δ_{AC} the antenna cable delay, and δ_R the receiver delay, that is, between the antenna cable

connector and the internal reference where the measurement is made. These instrumental delays are present in the term errors of (41.4) When extracting them explicitly one gets

$$P = \|\mathbf{x}_s - \mathbf{x}_r\| + c(\Delta t_{\text{rec}} - \Delta t_{\text{sat}}) + B(\text{rec}) + \epsilon, \quad (41.9)$$

where $B(\text{rec})$ is the bias associated with the signal delay across the antenna, the antenna cable, and the receiver

$$B(\text{rec}) = \delta_A + \delta_{AC} + \delta_R. \quad (41.10)$$

This bias should therefore be removed from the code measurements to retrieve the accurate synchronization error ($t_{\text{rec}} - t_{\text{ref}}$) between the receiver clock and the reference time scale. $B(\text{rec})$ should be constant. It is however sensitive to temperature variations, so that a temperature stabilization is recommended in the receiver room, as well as choosing an antenna cable with low sensitivity to temperature variations.

The second category of delays consists of the synchronization error between the internal timing reference and the external clock to be examined. For receivers with an internal or external time interval counter (i.e., R1 and R2), the synchronization error is measured by the TIC. This measurement must however be corrected for:

- The delays in the cables and electronic devices transporting the 1 pps signal from the clock to the TIC, that is, $(\delta_{iC} + \delta_{CC})$ for receivers R1 or δ_{CC} for receivers R2.
- The delays in the cables and electronic devices transporting the 1 pps signal from the receiver clock to the TIC, that is, δ_0 .
- The synchronization error between the internal reference and the 1 pps output of the receiver, that is, δ_{iR} , to be provided by the manufacturer.

When the receiver clock is directly using the frequency and time signals from the external clock (i.e., R3), only the clock cable delay δ_{CC} should be measured, and added to the bias δ_{iR} provided by the manufacturer. This second category of delay has to be added from the GNSS solution to go back from the receiver clock to the external clock.

Finally the synchronization error between the clock and the reference of the satellite clock products is for R1 and R2

$$\begin{aligned} (T - t_{\text{ref}}) &= (t_{\text{rec}} - t_{\text{ref}})_{\text{PR}} \\ &\quad - (\delta_A + \delta_{AC} + \delta_R) \\ &\quad + \text{TIC} + \delta_{CC} + \delta_{iC} - \delta_{iR} - \delta_0, \end{aligned} \quad (41.11)$$

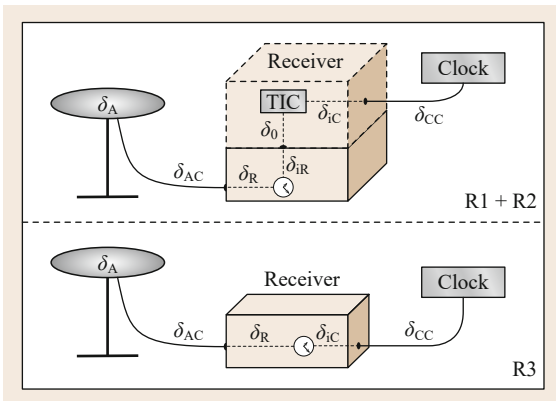


Fig. 41.11 Hardware delays to be accounted for time transfer and associated with three types of receivers of Fig. 41.10

while for the receiver R3 it reads

$$(T - t_{\text{ref}}) = (t_{\text{rec}} - t_{\text{ref}})_{\text{PR}} - (\delta_A + \delta_{\text{AC}} + \delta_R) + (\delta_{\text{CC}} + \delta_{\text{IC}}), \quad (41.12)$$

where $(\dots)_{\text{PR}}$ denotes the time offset as derived from pseudorange measurements.

The cable delays are independent of the signal frequency, and correspond to the product between the cable length and the group velocity of the signal in the cable given by

$$v_g = \frac{c}{\sqrt{\epsilon_r}}, \quad (41.13)$$

where v_g is the group velocity, c the velocity of light and ϵ_r the relative dielectric permittivity (Sect. 6.1). The connectors at both ends of the cable also induce some delays that must be taken into account. The group delay in the cable with its connectors can be measured with an accuracy of some tens of picoseconds using a time interval counter or a vector network analyzer (VNA). The offset δ_{IC} can only be determined if either the receiver offers an access to its internal clock via, for example, a 1 pps output synchronized on the internal clock, or the receiver manufacturer provides a detailed explanation of how the internal clock is constructed from the combination of input 1 pps and frequency coming from the external clock, so that it can be reproduced to be measured outside the receiver.

The antenna and receiver delays affecting the GNSS signals (δ_A and δ_R) are both frequency dependent. Two techniques exist to date for their calibration: the relative technique, using true GNSS signals, and the absolute technique using simulated GNSS signals. Note that the antenna calibration concerned here differs from the one discussed in 17.6.2 of this Handbook, which aims at determining the accurate phase center, while in the present case it determines the electric delay of the signal in the antenna. Neither elevation nor azimuth dependence of this delay is considered to date.

Absolute Calibration

The principle of the absolute calibration is to use simulated signals in order to determine the electric delay of the receiver or antenna (or the complete receiving chain), and compare the receiver/antenna measurements to the simulated signals. Complete descriptions of the method can be found, for example, in [41.38–40] and references therein. The simulated signal is produced by a GNSS signal generator, and is free of noise or perturbation like atmospheric delays or multipath existing in the case of true GNSS signals.

This kind of calibration offers a very high accuracy of 0.4 ns [41.41] while it requires the use of a GNSS simulator and a VNA, as well as an anechoic chamber for the antenna, which are not existing in the major part of the laboratories. Moreover, this method does not allow one to determine the hardware delays of already operational receiving chains, as these may not be interrupted and the antenna is calibrated in nonreal conditions.

Relative Calibration

Much more simple to be technically implemented is the relative calibration technique, which is consequently used for all operational stations. It consists in a comparison of pseudorange measurements collected by the local receiving chain and a reference receiving chain traveling from laboratory to laboratory [41.42]. For this, both stations should be connected to the same clock and installed in co-location (Fig. 41.12), so that all the perturbations except the multipath are equal. The difference of the pseudoranges measured by the two receiving chains then contain only a difference in antenna position plus the differences between the hardware delays of the two stations. Using the same nomenclature as in Fig. 41.11, this gives for a given satellite s and a given code c

$$\begin{aligned} P_{\text{lab}}(c, s) - P_{\text{ref}}(c, s) &= \|\mathbf{x}_s - \mathbf{x}_{\text{lab}}\| - \|\mathbf{x}_s - \mathbf{x}_{\text{ref}}\| \\ &+ (\delta R + \delta A)_{\text{lab}} - (\delta R + \delta A)_{\text{ref}} \\ &+ (\delta_{\text{AC}} - \delta_{\text{CC}} - \delta_{\text{IC}})_{\text{lab}} \\ &- (\delta_{\text{AC}} - \delta_{\text{CC}} - \delta_{\text{IC}})_{\text{ref}} + \epsilon, \end{aligned} \quad (41.14)$$

where ϵ is the combined noise and multipath of the two stations. The terms $(\delta_{\text{AC}} - \delta_{\text{CC}} - \delta_{\text{IC}})$ can be measured and deduced from the receiver manufacturer information for both receivers. The difference of pseudoranges therefore provides the $(\delta R + \delta A)$ of the laboratory receiver chain with respect to the reference chain. If this reference chain has been absolutely calibrated, the relative calibration then gives access to the true hardware

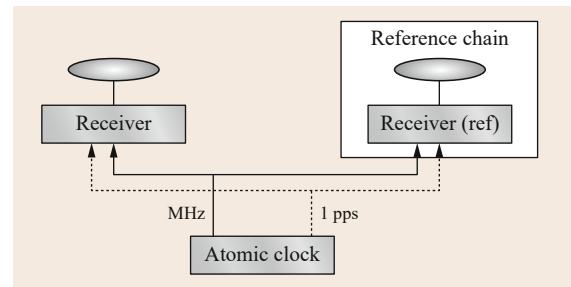


Fig. 41.12 Setup for relative calibration exercise

delay of the laboratory receiver+antenna. If not, then the relative calibration data can be used to calibrate a time transfer link in which the same reference chain was used for the calibration of both stations. In that case [41.43], the reference station has to be installed in co-location with the two stations of the link, and the calibration exercise provides the quantity $(\delta R + \delta A)_1 - (\delta R + \delta A)_2$ which can be directly applied to the time transfer solution $T_1 - T_2$. The same strategy can be applied to a network of stations in which all the stations are differentially calibrated with respect to a same reference. Any time link of the network will then be correctly calibrated by applying the computed relative hardware delays $(\delta R + \delta A)$ to each of the stations.

The relative calibration technique just described does not allow separating the hardware delays of the antenna and of the receiver. The isolated effect of receiver and antenna could be determined connecting the two receivers to the same antenna, using a splitter. However, the hardware delays of GNSS signals in the splitter are really difficult to be measured, and the splitter introduces a source of signal reflections, also called cable multipath, possibly inducing interferences [41.44]. Solutions to overcome that problem exist [41.40] but require the use of amplifiers and attenuators of which the levels have to be chosen thoroughly as a function of the antenna and receiver types. It is therefore recommended to use relative calibration only for the determination of the combined receiver plus antenna hardware delays.

The uncertainty budget of the differential calibration technique gives 2.3 ns for each isolated code [41.45], when taking into account the uncertainty on the absolute calibration of the reference chain, on the cable delay measurements, and the noise of the code measurements. In parallel, the uncertainty on the difference between two codes (e.g., L1/L2 P(Y) for GPS) is estimated to 2.0 ns so that the associated uncertainty

on the ionosphere-free combination is 3.8 ns. Considering a time transfer between two stations independently calibrated provides a type B uncertainty on the link at the level of 5.4 ns reduced to 5 ns in BIPM circular T. However, this 5 ns uncertainty reflects the long-lasting conservative practice. A significantly reduced uncertainty was found in [41.43] using a traveling receiver. The technique was then refined to reach an uncertainty around 1 ns in [41.46] and [41.47]. Such small u_B value however can be maintained over long times only if periodic re-calibrations are made.

The challenge for relative calibration is furthermore to keep constant the hardware delays of the reference station. The reference equipment is always subject to possible damages or instabilities due to its traveling between stations. The local temperature and humidity conditions in different locations can furthermore be very different from the conditions during the absolute (or relative) calibration of the reference equipment, which can cause some biases in the results [41.48]. While receivers can be installed in temperature-controlled rooms, the antenna and antenna cables can suffer diurnal temperature changes of around 40 °C, as occur in certain parts of the world. Some experiments of measuring the temperature sensitivities of the antennas showed maximum diurnal variations (for diurnal variations of 20 °C) of 40 ps for the carrier-phases [41.49], while up to 2 ns for the code measurements [41.50, 51]. Special attention should therefore be paid to sensitivity to temperature variations for the choice of the antenna and cable of the traveling reference GNSS station, as well as any other GNSS station dedicated to time transfer. The stability of the reference GNSS station should furthermore be regularly verified, by intercomparison with fixed stations. The delays of the most stable GPS common-view time transfer receivers vary typically by a few nanoseconds over years, generally by less than 5 ns peak-to-peak [41.52].

41.4 Multi-GNSS Time Transfer

41.4.1 General Requirements

The combination of measurements from different GNSS constellations for time transfer requires several specifications. The first one is that the receiver internal reference be the same for all systems. A second requirement is that the receiver must be fully calibrated, that is, the hardware delays must be determined for each signal transmitted by each constellation. Indeed, in some cases the frequency bands used by different systems do not completely overlap, or the power spectrum inside

the band is not the same. For example in GLONASS, several carrier frequencies are used in each frequency bands (Chap. 8), yielding complex calibration procedures due to the need to calibrate one delay per carrier frequency.

Finally, a last requirement concerns the reference of the satellite clock products. The receiver clock solution obtained from observations of satellites belonging to constellations A and B are $(t_{\text{rec}} - t_{\text{ref,A}})$ and $(t_{\text{rec}} - t_{\text{ref,B}})$, where $t_{\text{ref,A}}$ and $t_{\text{ref,B}}$ are the reference time scales of the two constellations. In order to get only one combined

receiver clock solution, the user should either know the accurate de-synchronization ($t_{\text{ref,A}} - t_{\text{ref,B}}$) at each observation epoch, or use satellite clock products having the same reference whatever the constellation to which they belong, or introduce ($t_{\text{ref,A}} - t_{\text{ref,B}}$) as unknown and then estimate it along with the other parameters. As ($t_{\text{ref,A}} - t_{\text{ref,B}}$) is generally not available at each observation epoch, only the two other possibilities can be used. The last one however requires the estimation of one additional parameter at each observation epoch which increases the uncertainty of the solution. The optimal option is therefore the second one. Combined products already exist for GPS and GLONASS satellites, they are provided by some analysis centers of the IGS. It is assumed that in the future, such products will also be provided for Galileo and BeiDou. The combination of all these constellations in one global time transfer solution will therefore be possible with the second option. We describe here the case of GLONASS which requires a special treatment due to existence of interfrequency biases, and present some first results on the use of Galileo and BeiDou for time and frequency transfer. Note that QZSS can also be used for regional clock comparisons, while it will play no role in intercontinental time transfer.

41.4.2 GPS + GLONASS Combination

The main difference between GLONASS and the other GNSS is the channel access method. While all the GNSS constellations use the code division multiplex access (CDMA) technique in which all the satellites share the same carrier frequencies, GLONASS is based on the frequency division multiple access (FDMA) technique. Each GLONASS satellite transmits consequently on a different frequency in the L1 band as well as in the L2 band.

Due to the frequency-dependent nature of the hardware delays in the receiver and in the antenna, these are different for each GLONASS satellite group transmitting a given pair of frequencies L1, L2, inducing interfrequency biases up to tens of nanosecond in the code measurements and hence in the clock solutions as determined from different satellites. The observation equation (41.9) should therefore be modified for GLONASS satellites as

$$P = \|\mathbf{x}_s - \mathbf{x}_r\| + c(\Delta t_{\text{rec}} - \Delta t_{\text{sat}}) + B(\text{rec}, \text{sat}) + \epsilon. \quad (41.15)$$

In this case, the bias $B(\text{rec}, \text{sat})$ being satellite dependent cannot be absorbed by the receiver clock Δt_{rec} . Note that, in principle, bias should be the same for satellites using the same frequency pair, but it is rarely modeled

as such, and a satellite-dependent bias is generally preferred.

The satellite clock Δt_{sat} , the receiver clock Δt_{rec} , and the bias $B(\text{rec}, \text{sat})$ in equation (41.15) cannot be separated unequivocally. As a consequence, the GLONASS satellite clocks determined from some network analysis are affected by artificial biases. In that computation, it is indeed necessary to fix arbitrarily one bias for a given receiver–satellite pair and then to determine all satellite clocks, receiver clocks and receiver–satellite biases with respect to that fixed parameter. If the fixed bias changes between the treatments of two successive data batches, the biases for all station–satellite pairs will change accordingly. As classically the satellite clock products are computed on a daily basis, their use to determine the clock solution of a single station (in AV or PPP) requires the estimation of daily receiver–satellite biases $B'(\text{rec}, \text{sat}, \text{day})$ in addition to the clock solution. These biases contain a physical part, corresponding to the station hardware delays for the frequency emitted by the satellite, and which is constant (or nearly constant) over the long-term, plus an artificial bias present in the satellite clock products

$$B'(\text{rec}, \text{sat}, \text{day}) = B(\text{rec}, \text{sat}) + \gamma(\text{sat}, \text{day}), \quad (41.16)$$

$B(\text{rec}, \text{sat})$ corresponds to the terms ($\delta_A + \delta_{AC} + \delta_R$) in (41.11), that is, the sum of antenna delay, antenna cable delay and receiver delay.

As the biases $\gamma(\text{sat}, \text{day})$ are the same for all the GNSS stations, they cancel out in the common view approach (Sect. 41.2.3) and a calibrated GLONASS clock solution can be obtained if $B(\text{rec}, \text{sat})$ is known for both stations of the link, that is, if the stations have been calibrated for all the GLONASS frequencies.

However, such a calibrated clock solution cannot be obtained with the PPP and AV techniques using only GLONASS measurements due to the unknown biases $\gamma(\text{sat}, \text{day})$. The combination of GPS plus GLONASS measurements allows one to solve for that issue. Two approaches are then possible: the first one uses only the GPS calibration results and determines the biases $B'(\text{rec}, \text{sat}, \text{day})$ as the differences for each satellite and each day, between the noncalibrated GLONASS clock results and the calibrated GPS clock solution.

An application of this technique for the computation of all in view solutions based on CGGTTS results was presented in [41.53], and the corresponding combination of GPS and GLONASS in precise point positioning can be found in [41.54]. However, in both cases the GLONASS measurements are not calibrated which is not convenient from the metrological point of view but is a consequence of the unknown bias present in the satellite clock products. The second possibility is to use

a link approach. The basic idea is that the differences between the estimated $B'(\text{rec}, \text{sat}, \text{day})$ for the two stations of the link reads

$$\begin{aligned} B'(\text{rec}_1, \text{sat}, \text{day}) - B'(\text{rec}_2, \text{sat}, \text{day}) \\ = B(\text{rec}_1, \text{sat}) - B(\text{rec}_2, \text{sat}) \end{aligned} \quad (41.17)$$

that is, does not depend any more on the biases of the satellite clock products, and can be accurately determined by a calibration exercise. The values of $B(\text{rec}_1, \text{sat}) - B(\text{rec}_2, \text{sat})$ can then be used to constrain the determination of $B'(\text{rec}_1, \text{sat}, \text{day})$ and $B'(\text{rec}_2, \text{sat}, \text{day})$. This requires, of course, that the clock solutions (AV or PPP solutions) of the two stations are determined in a same analysis procedure [41.55]. Finally, the clock solutions obtained from the combination of GLONASS and GPS measurements provide a same level of performances than the GPS-only solutions, as shown in the publications cited earlier.

41.4.3 Time Transfer with Galileo and BeiDou

As stated in Chap. 9, Galileo is transmitting in three frequency bands (E1, E5, and E6), but only the two former ones are available in the open service. Most dual-frequency receivers measure the unencrypted ranging codes E1 and E5a, and improved-accuracy receivers measure additionally the signal E5b and the wide-band E5 alternative binary offset carrier (AltBOC) signal.

Some first experiments of using the Galileo signals for time transfer were already realized, based on the ionosphere-free combinations of E1 with either E5a, or E5b or E5 AltBOC. The results indicate that the

noise of the Galileo measurements is significantly lower than the noise of the ionosphere-free combination of the GPS measurements P(Y)-codes on L1 and L2 at all elevations [41.56]. This comes partly from the smaller coefficients multiplying the code measurements (and hence the noise) in the ionosphere-free combination. These coefficients are indeed smaller for more distant frequencies, as is for (L1, L5) with respect to (L1, L2). The noise of the final solution however depends of the number of visible satellites so that Galileo will compete with GPS only when the full constellation will be deployed.

First steps in BeiDou time transfer have also been started [41.57]. From theoretical point of view, the noise of the ionosphere-free combination should not be so much lower than the present GPS noise, due to the proximity of the frequencies of the open service (B1 and B2), but no rigorous comparison exists to date.

It must however be noted that as GPS, Galileo, and BeiDou are based on the CDMA technique, so that a same total hardware delay affects the code measurements from all the Galileo satellites, and a same corresponding hardware delay affects the code measurements from all the BeiDou satellites. No satellite-dependent hardware delay should be determined as is the case for GLONASS. The combination of GPS with Galileo and/or BeiDou will therefore increase the number of observations without increasing the number of unknowns (except in PPP where additional phase ambiguities will have to be resolved), and an improvement by a factor of $\sqrt{3}$ is expected from the combination of GPS with the full Galileo and BeiDou constellations.

41.5 Conclusions

This chapter presented the time and frequency applications offered by the GNSS, as well as their current performances summarized in Table 41.1. Accurate time or time transfer can be realized with an accuracy approaching one nanosecond thanks to most advanced calibration techniques. Due to the noise and multipath of the code pseudoranges, the accurate frequency transfer will be preferably determined with PPP which makes use of the carrier-phase measurements, and which allows comparing atomic clocks at the level of $1 \cdot 10^{-13}$ for short averaging times (some minutes) and approaching $1 \cdot 10^{-16}$ at one day averaging times. These ultimate performances can of course be reached only with appropriate instrumentation.

In the future, GNSS may grow to include more than 100 satellites, mostly in medium Earth orbit, with some in geostationary and inclined elliptical orbits. Using the

Table 41.1 Summary of the best performances of GNSS timing applications

Application	Parameter	Performance
Synchronization on UTC ^a	μ_B	20 ns
	μ_A	10 ns
Frequency steering (GNSSDO) ^b	Stability	$\leq 1 \cdot 10^{-12}$ one day
Time transfer with CGGTTS ^c	μ_B	< 2 ns
	μ_A	2 ns
Time transfer with PPP ^c	μ_B	< 2 ns
	μ_A	100 ps
Frequency transfer with PPP ^b	Stability	$2 \cdot 10^{-16}$ one day

^a Function of the GNSS prediction of UTC and of the receiver calibration

^b Given by the Allan deviation

^c Function of the receiver calibration

same kind of signals from different constellations will increase the number of measurements in the averaging procedure, and hence produce a slight improvement in terms of precision [41.58]. Furthermore, each of the GNSSs will be gradually modernized with, for example, new onboard clocks for GPS, or transfer to the CDMA technique for GLONASS which will eliminate the inter-channel biases.

However, the added value of these constellations resides also in the new possibilities they offer thanks to new ranging signals having more complex structure and improved characteristics. We can, for example, expect new timing performances thanks to the precise Galileo E5 AltBOC signal, whose combined noise and multipath is limited to less than 25 cm at all satellite elevations [41.59]. If in the future, GNSSs offer some new signals in the C-band (or Ku band), the ionosphere-free combination with code measurements in the L-band would also mitigate the noise amplification due to the coefficients of the linear combination.

In parallel, some new clock comparison techniques will be offered by the GNSS for fundamental time metrology. BeiDou, for example, will provide, besides

the timing services described in this chapter, a two-way time transfer, that is, based on signals that travel both ways between the two ground clocks to be compared via a BeiDou geostationary satellite equipped with a transponder [41.60]. This technique is already widely used by the time laboratories for their participation to TAI [41.61] but relies presently on commercial satellites. BeiDou will be the first navigation system to propose that alternative to the classical one-way method. The station equipment for that technique will however be completely different from the classical GNSS receiving station, as an emission unit must be considered as well. BeiDou and GLONASS will additionally allow laser time transfer as the satellites are equipped with laser reflectors, offering to compare remote ground clocks with accuracy levels not achievable by radio systems. A first experiment with BeiDou demonstrated a precision of approximately 300 ps on the clock difference and relative frequency stability of $1 \cdot 10^{-14}$ for the comparison between a ground hydrogen maser and satellite rubidium clocks [41.62]. Finally, future generations of the GNSS, still under study, will open in the next decades still new horizons for time and frequency metrology.

References

- 41.1 H.-G. Berns, T.H. Burnett, R. Gran, R.J. Wilkes: GPS time synchronization in school-network cosmic ray detectors, *IEEE Trans. Nucl. Sci.* **51**(3), 848–853 (2004)
- 41.2 E. Butterline, J. Abate, G. Zampetti: Use of GPS to synchronize the AT&T national telecommunications network, *Proc. 21th Annu. PTTI Appl. Plan. Meet.* (1988) pp. 65–75
- 41.3 I. Hall, P.G. Beaumont, P.G. Baber, I. Shuto, M. Saga, K. Okuno, H. Uo: New line current differential relay using GPS synchronization, *Proc. IEEE Power Tech Conf., Bologna* (2003) pp. 1–8
- 41.4 Joint Committee for Guides in Metrology: Evaluation of measurement data: Guide to the expression of uncertainty in measurement, *JCGM 100:2008*, http://www.bipm.org/utls/common/documents/jcgm/JCGM_100_2008_E.pdf (2008)
- 41.5 E.F. Arias: The metrology of time, *Phil. Trans. R. Soc. A* **363**, 2289–2305 (2005)
- 41.6 M. Lombardi: The use of GPS disciplined oscillators as primary frequency standards for calibration and metrology laboratories, *Measure* **3**(3), 56–65 (2008)
- 41.7 The OPERA Collaboration, T. Adam, N. Agafonova, A. Aleksandrov, O. Altinok, P. Alvarez Sanchez, A. Anokhina, S. Aoki, A. Ariga, T. Ariga, D. Autiero, A. Badertscher, A. Ben Dhahbi, A. Bertolin, C. Bozza, T. Brugière, R. Brugnera, F. Brunet, G. Brunetti, S. Buontempo, B. Carlus, F. Cavanna, A. Cazes, L. Chaussard, M. Chernyavsky, V. Chiarella, A. Chukanov, G. Colosimo, M. Crespi, N. D'Ambrosio, G. De Lellis, M. De Serio, Y. Déclais, P. del Amo Sanchez, F. Di Capua, A. Di Crescenzo, D. Di Ferdinando, N. Di Marco, S. Dmitrievsky, M. Dracos, D. Duchesneau, S. Dusini, T. Dzhatdov, J. Ebert, I. Efthymiopoulos, O. Egorov, A. Ereditato, L.S. Esposito, J. Favier, T. Ferber, R.A. Fini, T. Fukuda, A. Garfagnini, G. Giacomelli, M. Giorgini, M. Giovannozzi, C. Gierd, J. Goldberg, C. Göllnitz, D. Golubkov, L. Goncharova, Y. Gornushkin, G. Grella, F. Grianti, E. Gschwendtner, C. Guerin, A.M. Guler, C. Gustavino, C. Hagner, K. Hamada, T. Hara, R. Enikeev, M. Hierholzer, A. Hollnagel, M. Ieva, H. Ishida, K. Ishiguro, K. Jakovcic, C. Jollet, M. Jones, F. Juget, M. Kamiscioglu, J. Kawada, S.H. Kim, M. Kimura, E. Kiritis, N. Kitagawa, B. Klicek, J. Knuesel, K. Kodama, M. Komatsu, U. Kose, I. Kreslo, C. Lazzaro, J. Lenkeit, A. Ljubicic, A. Longhin, A. Malgin, G. Mandrioli, J. Marteau, T. Matsuo, V. Matveev, N. Mauri, A. Mazzoni, E. Medinaceli, F. Meisel, A. Mereaglia, P. Migliozi, S. Mikado, D. Missiaen, P. Monacelli, K. Morishima, U. Moser, M.T. Muciaccia, N. Naganawa, T. Naka, M. Nakamura, T. Nakano, Y. Nakatsuka, D. Naumov, V. Nikitina, F. Nitti, S. Ogawa, N. Okateva, A. Olchevsky, O. Palamara, A. Paoloni, B.D. Park, I.G. Park, A. Pastore, L. Patrizii, E. Pennacchio, H. Pessard, C. Pistillo, N. Polukhina, M. Pozzato, K. Pretzl, F. Pupilli, R. Rescigno, F. Riguzzi, T. Roganova, H. Rokujo, G. Rosa, I. Rostovtseva, A. Rubbia, A. Russo, V. Rysany, O. Ryazhskaya, O. Sato, Y. Sato, Z. Sah-

- noun, A. Schembri, J. Schuler, L. Scotto Lavina, J. Serrano, I. Shakiryanova, A. Sheshukov, H. Shibuya, G. Shoziyoev, S. Simone, M. Sioli, C. Sirignano, G. Sirri, J.S. Song, M. Spinetti, L. Stanco, N. Starkov, S. Stellacci, M. Stipcevic, T. Strauss, S. Takahashi, M. Tenti, F. Terranova, I. Tezuka, V. Tioukov, P. Tolun, N.T. Trani, S. Tufanli, P. Vilain, M. Vladimirov, L. Votano, J.-L. Vuilleumier, G. Wilquet, B. Wonsak, J. Wurtz, V. Yakushev, C.S. Yoon, J. Yoshida, Y. Zaitsev, S. Zemskova, A. Zghiche: Measurement of the neutrino velocity with the OPERA detector in the CNGS beam, *J. High Energy Phys.* **2012**(10), 1–37 (2012)
- 41.8 D. Mills: *Computer Network Time Synchronization: The Network Time Protocol on Earth and in Space*, 2nd edn. (CRC, Boca Raton 2012)
- 41.9 D.W. Allan, M. Weiss: Accurate time and frequency transfer during common-view of a GPS satellite, *Proc. IEEE FCS 1980, Philadelphia* (1980) pp. 334–356
- 41.10 G. Petit: The TAI PPP pilot experiment, *Proc. Joint IEEE FCS and 23rd EFTF, Besançon* (2009) pp. 116–119
- 41.11 P. Defraigne, G. Petit: CGGTS-Version 2E: An extended standard for GNSS time transfer, *Metrologia* **52**(6), G1 (2015)
- 41.12 D.W. Allan, C. Thomas: Technical directives for standardization of GPS time receiver software, *Metrologia* **31**, 69–79 (1994)
- 41.13 J. Levine: Time transfer using multi-channel GPS receivers, *IEEE Trans. Ultrason. Ferroelectr. Freq. Contr.* **46**(2), 284–291 (1999)
- 41.14 M. Hernández-Pajares, J.M. Juan, J. Sanz, R. Orus, A. Garcia-Rigo, J. Feltens, A. Komjathy, S.C. Schaer, A. Krankowski: The IGS VTEC maps: A reliable source of ionospheric information since 1998, *J. Geod.* **83**, 263–275 (2009)
- 41.15 P. Defraigne, G. Petit: Time transfer to TAI using geodetic receivers, *Metrologia* **40**, 184–188 (2003)
- 41.16 G. Petit, Z. Jiang: GPS All in view time transfer for TAI computation, *Metrologia* **45**, 35–45 (2008)
- 41.17 M.A. Weiss, G. Petit, Z. Jiang: A comparison of GPS common-view time transfer to all-in-view, *Proc. IEEE FCS, Vancouver* (2005) pp. 1–5
- 41.18 M.C. Martínez-Belda, P. Defraigne: Combination of TWSTFT and GPS data for time transfer, *Metrologia* **47**, 305–316 (2010)
- 41.19 T. Schildknecht, G. Beutler, M. Rothacher: Towards sub-nanosecond GPS time transfer using geodetic processing technique, *Proc. 4th EFTF, Neuchâtel* (1990) pp. 335–346
- 41.20 K.M. Larson, J. Levine, L.M. Nelson, T. Parker: Assessment of GPS carrier-phase stability for time-transfer applications, *IEEE Trans. Ultrason. Ferroelectr. Freq. Contr.* **47**(2), 484–494 (2000)
- 41.21 C. Bruyninx, P. Defraigne: Frequency transfer using GPS codes and phases: Short and long term stability, *Proc. 31st PTI Meet., Dana Point*, ed. by L.A. Breakiron (USNO, Washington DC 2000) pp. 471–478
- 41.22 K. Senior, J. Ray: Accuracy and precision of carrier phase clock estimates, *Proc. 33rd PTI Meet., Long Beach*, ed. by L.A. Breakiron (USNO, Washington DC 2001) pp. 199–217
- 41.23 J. Ray, K. Senior: Geodetic techniques for time and frequency comparisons using GPS phase and code measurements, *Metrologia* **42**(4), 215–232 (2005)
- 41.24 P. Defraigne, C. Bruyninx: Multipath mitigation in GPS-based time and frequency transfer, *Proc. 20th EFTF, Braunschweig* (2006) pp. 524–529
- 41.25 D. Orgiazzi, P. Tavella, F. Lahaye: Experimental assessment of the time transfer capability of precise point positioning (PPP), *Proc. IEEE FCS, Vancouver* (2005) pp. 337–345
- 41.26 K. Senior, E. Powers, D. Matsakis: Attenuating day-boundary discontinuities in GPS carrier-phase time transfer, *Proc. 31st Precise Time Interval Syst. Appl. (PTI) Meet., Dana Point* (USNO, Washington DC 2000) pp. 481–490
- 41.27 N. Guyennon, G. Cerretto, P. Tavella, F. Lahaye: *Further characterization of the time transfer capabilities of precise point positioning (PPP)*, *PROCBE-GINProc. Joint IEEE Freq. Contr. Symp. 21st Eur. Freq. Time Forum, Geneva* (2007) pp. 399–404
- 41.28 J. Delporte, F. Mercier, D. Laurichesse: Time transfer using GPS carrier phase with zero-difference integer ambiguity blocking, *Proc. 22nd EFTF, Toulouse* (2008) pp. 1–6
- 41.29 G. Petit, A. Harmegnies, F. Mercier, F. Perosanz, S. Loyer: The time stability of PPP links for TAI, *Proc. Joint IEEE FCS and 25th EFTF, San Francisco* (2011) pp. 1–5
- 41.30 F. Lahaye, P. Collins, G. Cerretto, P. Tavella: Advances in time and frequency transfer from dual-frequency GPS pseudorange and carrier-phase observations, *Proc. 40th PTI Syst. Appl. Meet., Reston* (2009) pp. 415–432
- 41.31 J. Ray: Systematic errors in GPS position estimates. Presentation at IGS 2006 Workshop, Darmstadt (available electronically at http://figs.bjpl.nasa.gov/pub/resource/pubs/06_darmstadt/IGS%20Presentations%20PDF/11_6_Ray.pdf)
- 41.32 W. Aerts, Q. Baire, C. Bruyninx, J. Legrand, E. Pottiaux: Towards better GNSS observations at the new IGS reference station BRUX: Multi path mitigation and individual antenna calibration, *Proc. AGU Fall Meet., San Francisco* (AGU, Washington 2012), abstract No. G51C-07
- 41.33 Q. Baire, P. Defraigne, E. Pottiaux: Influence of troposphere in PPP time transfer, *Proc. Joint IEEE FCS and 23rd EFTF, Besançon* (2009) pp. 1065–1068
- 41.34 U. Weinbach, S. Schön: On the correlation of tropospheric zenith path delay and station clock estimates in geodetic GNSS frequency transfer, *Proc. 24th Eur. Freq. Time Forum, Noordwijk* (2010) pp. 1–8
- 41.35 P. Defraigne, G. Petit, C. Bruyninx: Use of geodetic receivers for TAI, *Proc. 31st PTI Syst. Appl. Meet., Long Beach*, ed. by L.A. Breakiron (USNO, Washington DC 2002) pp. 341–348
- 41.36 Bureau International des Poids et Mesures: Consultative Committee for Time and Frequency (CCTF) <http://www.bipm.org/utls/en/pdf/CCTF15-EN.pdf>
- 41.37 P. Defraigne, G. Petit, P. Uhrich, W. Aerts: Requirements on GNSS receivers from the perspective of timing applications, *Proc. 24th Eur. Freq. Time Fo-*

- rum, Noordwijk (2010) pp. 1–6
- 41.38 J. White, R. Beard, G. Landis, G. Petit, E. Powers: Dual frequency absolute calibration of a geodetic GPS receiver for time transfer, Proc. 15th EFTF, Neuchâtel (2001) pp. 167–172
- 41.39 G. Cibieli, A. Proia, L. Yaigre, J.-F. Dutrey, A. de La-tour, J. Dantepal: Absolute calibration of geodetic receivers for time transfer: Electrical delay measurement, uncertainties and sensitivities, Proc. 22nd EFTF, Toulouse (CNES, Toulouse 2008) pp. 1–7
- 41.40 J. Plumb, K. Larson, J. White, E. Powers: Absolute calibration of a geodetic time transfer system, IEEE Trans. Ultrason. Ferroelectr. Freq. Contr. **52**(11), 1904–1911 (2005)
- 41.41 A. Proia, G. Cibieli, L. Yaigre: Time stability and electrical delay comparison of dual frequency GPS receivers, Proc. 44th Annu. PTI Meet. (2012) pp. 297–302
- 41.42 G. Petit, Z. Jiang, P. Uhrich, F. Taris: Differential calibration of Ashtech Z12-T receivers for accurate time comparisons, Proc. 14th EFTF, Torino (Swiss Foundation for Research in Microtechnology, Neuchâtel 2000) pp. 40–44
- 41.43 H. Esteban, J. Palacio, F.J. Galindo, T. Feldmann, A. Bauch, D. Piester: Improved GPS-based time link calibration involving ROA and PTB, IEEE Trans. Ultrason. Ferroelectr. Freq. Contr. **57**(3), 714–720 (2010)
- 41.44 M. Weiss, F. Ascarrunz, T. Parker, V. Zhang, X. Gao: Effects of antenna cables on GPS timing receivers, Proc. Joint IEEE FCS and 13th EFTF, Besançon (1999) pp. 259–262
- 41.45 G. Petit, P. Defraigne, B. Warrington, P. Uhrich: Calibration of dual frequency GPS receivers for TAI, Proc. 20th EFTF, Braunschweig (PTB, Braunschweig 2006) pp. 455–459
- 41.46 T. Feldmann, A. Bauch, D. Piester, M. Rost, E. Goldberg, S. Mitchell, B. Fonville: Advanced GPS-based time link calibration with PTB's new GPS calibration setup, Proc. 42nd PTI Syst. Appl. Meet., Reston (2011) pp. 509–526
- 41.47 D. Rovera, J.-M. Torre, R. Sherwood, M. Abgrall, C. Courde, M. Laas-Bourez, P. Uhrich: Link calibration against receiver calibration: An assessment of GPS time transfer uncertainties, Metrologia **51**(5), 476–490 (2014)
- 41.48 S.F. Adam: *Microwave Theory and Applications*, 2nd edn. (Prentice Hall, Upper Saddle River 1969)
- 41.49 J. Ray, K. Senior: Temperature sensitivity of timing measurements using Dorne Margolin antennas, GPS Solutions **2**(1), 24–30 (2001)
- 41.50 A. Smolarsk, A. Lisowiec, J. Nawrocki: Improving the accuracy of GPS time transfer by thermal stabilization of GPS antenna and receiver, Proc. 16th EFTF, St. Petersburg (Swiss Foundation for Research in Microtechnology, Neuchâtel 2002) pp. 503–505
- 41.51 P. Defraigne, C. Bruyninx: On the link between GPS pseudorange noise and day-boundary discontinuities in geodetic time transfer solutions, GPS Solutions **11**(4), 239–249 (2007)
- 41.52 M. Weiss, W. Lewandowski, P. Uhrich, D. Valat: NIST and OP GPS receiver calibrations spanning twenty years: 1983–2003, Proc. 18th EFTF, Guilford (Univ. of Surrey, Surrey 2004) pp. 143–146
- 41.53 A. Harmegnies, P. Defraigne, G. Petit: Combining GPS and GLONASS in all-in-view for time transfer, Metrologia **50**(3), 277–287 (2013)
- 41.54 P. Defraigne, Q. Baire: Combining GPS and GLONASS for time and frequency transfer, Adv. Space Res. **47**(2), 265–275 (2011)
- 41.55 P. Defraigne, W. Aerts, A. Harmegnies, G. Petit, D. Rovera, P. Uhrich: Advances in multi-GNSS time transfer, Proc. Joint IEEE FCS and EFTF 2013, Prague (2013) pp. 508–512
- 41.56 P. Defraigne, W. Aerts, G. Cerretto, G. Signorile, E. Cantoni, I. Sesia, P. Tavella, A. Cernigliaro, A. Samperi, J.M. Sleewaegen: Advances on the use of Galileo signals in time metrology: Calibrated time transfer and estimation of UTC and GGT0 using a combined commercial GPS-Galileo receiver, Proc. PTI Syst. Appl. Meet. (2014) pp. 256–262
- 41.57 W. Guang, H. Yuan: The application of smoothed code in BeiDou common view, Proc. CSNC 2013, Wuhan, Vol. I, ed. by J. Sun, W. Jiao, H. Wu, C. Shi (Springer, Berlin 2013) pp. 269–278
- 41.58 J. Furthner, A. Moudrak, A. Konovaltsev, J. Hammesfahr, H. Denks: Time dissemination and common view time transfer with Galileo: How accurate will it be?, Proc. 35th Annu. PTI Meet., San Diego (2004) pp. 185–198
- 41.59 A. Simsky, D. Mertens, J.M. Sleewaegen, W. De Wilde, S. Navigation, M. Hollreiser: Multipath and tracking performance of Galileo ranging signals transmitted by GIOVE-B, Proc. ION GNSS 2008, Savannah (ION, Virginia 2008) pp. 1525–1536
- 41.60 W.K. Yang, H. Gong, Z.J. Liu, Y.L. Li, G.F. Sun: Improved two-way satellite time and frequency transfer with multi-GE0 in BeiDou navigation system, Sci. China Inf. Sci. **57**(2), 1–15 (2014)
- 41.61 A. Bauch, J. Achkar, S. Bize, D. Calonico, R. Dach, R. Hlavac, L. Lorini, T. Parker, G. Petit, D. Piester, K. Szymaniec, P. Uhrich: Comparison between frequency standards in Europe and the USA at the 1015 uncertainty level, Metrologia **43**, 109–120 (2006)
- 41.62 W. Meng, H. Zhang, P. Huang, J. Wang, Z. Zhang, Y. Liao, Y. Ye, W. Hu, Y. Wang, W. Chen: Design and experiment of onboard laser time transfer in Chinese Beidou navigation satellites, Adv. Space Res. **51**(6), 951–958 (2013)

Annex A: Data Formats

Oliver Montenbruck, Ken MacLeod

GNSS formats have developed within government, industry, and academia. Their standardization has facilitated the efficient development of the GNSS industry. Current GNSS formats support meta-data such as GNSS station, receiver, antenna and equipment calibration information, GNSS observation and broadcast navigation information and also GNSS products such as precise orbits, clock corrections, atmospheric measurements and station coordinates. RINEX, BINEX and IGS standards have been widely accepted and have become de facto standards. RTCM de jure standards on the other hand have developed within a standards organization and were adopted by industry. The development of new GNSS constellations has led to the need for new formats and also encouraged a higher level of cooperation and integration between the GNSS standards groups. The most widely used GNSS standards are described in this chapter.

The GNSS community relies on a variety of standards that have been developed by different entities to facilitate an interoperable and efficient exchange of data and products between providers and users.

Even though most manufacturers employ company specific – and sometimes nondisclosed – message formats for communication with their GNSS receivers, a variety of nonvendor-specific formats have been developed by various nonprofit organizations [A.1]:

- RTCM SC-104 standard for Differential GNSS Services of the Radio Technical Commission for Maritime Services (RTCM)
- The GNSS-related parts of the NMEA 0183 interface standard of the National Marine Electronics Association (NMEA)
- The Receiver INdependent EXchange (RINEX) format developed by the International GNSS Service (IGS and RTCM SC-104)
- The BINary EXchange (BINEX) format of the University NAVSTAR consortium (UNAVCO).

A subset of these protocols and formats is typically supported in addition (or alternative) to proprietary data formats by all receivers. However, the various standards coexist with each other and the specific standard(s) sup-

ported by a given receiver depends largely on the type of user and application.

Complementary to the aforementioned standards for the exchange of GNSS receiver data, a variety of standards have been developed by the IGS to harmonize the exchange of products and metadata. Examples include:

- The Standard Product 3 (SP3) for orbit and clock information,
- The Clock RINEX format for the exchange of postprocessed satellite and receiver clock offset solutions,
- The IONosphere EXchange format (IONEX),
- The ANTenna EXchange (ANTEX) format for the provision of antenna phase center offsets and variations,
- The SiteLog format for station related information, and
- The Solution INdependent EXchange (SINEX) format for a harmonized exchange of estimated parameter sets.

In view of their primary application within the IGS community, the latter standards are generally more open and flexible than formal standards established by industrial or governmental standardization organizations.

Within the following sections the key features of the various standards are briefly described along with illustrating examples. For a more detailed discussion and concise definitions the readers are referred to the official documentation published by the corresponding organization.

A.1 Receiver Formats

A.1.1 NMEA 0183

The National Marine Electronics Association (NMEA) [A.2] is a nonprofit organization that was founded in 1957 by a group of electronic dealers with the goal of strengthening their relationships with electronic manufacturers. NMEA standards include the NMEA 0183 standard (using an ASCII text format) and a modernized version NMEA 2000 (binary). Both standards enable interoperability between marine electronics and support communications as well as data message standards.

The legacy standard (NMEA 0183 [A.3]), which has widely been adopted in GNSS receivers for marine, aeronautical and personal navigation, provides generic specifications of electrical requirements, protocols and message formats for a wide range of devices via a serial interface bus. Such devices may include electronic chart display and information systems, timekeeping equipment, radars, heading sensors, and sounders, LORAN-C (LOng RANge Navigation) receivers, as well as receivers for global navigation satellite systems.

All NMEA 0183 data are transferred in the form of ASCII text messages via an RS422 (or, alternatively, RS232) serial interface. At a nominal rate of 4800 baud, approximately 600 characters can be transmitted per second, which is generally compatible with the needs of communicating position, speed and auxiliary data of slow moving vessels. Each device sending data is assigned a two letter *talker ID* reflecting the equipment type (Table A.1).

All messages start with a \$ sign followed by a five-character string comprising the talker ID (tt) and a three character message ID (mmm):

```
$ttmmm,d1,d2,...*hh<CR><LF>
```

Following the message header, individual data fields d1,d2,... made up of numbers or characters are provided as a comma separated list. The sequence of parameters is specific for each message type. Optional values may be omitted but the separating comma must be provided to enable an unambiguous decoding. The message is terminated by an optional, two-character checksum (introduced by an asterisk) and a set of carriage-return line-feed characters.

An overview of commonly used NMEA 0183 messages for GPS devices is provided in Table A.2. In addition to these, vendor-specific messages are supported by the standard and are in use by various receiver manufacturers. These are identified by \$P and a three-letter vendor ID instead of the standard message header (e.g., \$PUBX and \$PASH for uBlox and Ashtech receivers), but otherwise follow the message concept described above.

As an example, the contents of the most widely employed GPS position message (\$GPGGA) is illustrated in Fig. A.1. It provides the measured position in terms of geographic longitude and latitude along with the height above sea level but also the difference between ellipsoidal and geoid height assumed used in its computation. Since only the time-of-day is given in the \$GPGGA message, a complementary \$GPZDA date and time message may need to be output for a unique specification of the current epoch.

Even though common users are mostly interested in position-, velocity- and time-related information, the

Table A.1 Short list of selected NMEA 0183 talker identifications

ID	Device
EC	Electronic chart display & information system
GP	Global Positioning System receiver
IN	Integrated navigation system
RA	Radar
WI	Weather instrument
ZQ	Timekeeper (quartz)

Table A.2 Common NMEA 0183 GNSS messages

Msg ID	Description
\$GPALM	GPS almanac data
\$GPBOD	Bearing: origin to destination (UTC time, current position, true & magnetic bearing and distance to destination)
\$GPDTM	Datum reference
\$GPGGA	Global Positioning System fix data (UTC time, position, quality indicators, use of differential corrections)
\$GPGLL	Geographic position latitude/longitude
\$GPGRS	GPS range residuals
\$GPGSA	GPS DOP and active satellites (positioning mode, fix type, satellites used in fix, PDOP, HDOP and VDOP)
\$GPGST	GPS pseudorange noise statistics
\$GPGSV	GPS satellites in view (number satellites in view, satellite number, azimuth and elevation, signal-to-noise ratio)
\$GPHDT	True heading
\$GPRMC	Recommended minimum navigation information (UTC time and date, latitude and longitude, speed over ground, status, etc.)
\$GPVTG	Course over ground and ground speed
\$GPZDA	Time and date (UTC time, calendar date, local time zone UTC offset)

NMEA 0183 standard also supports a variety of lower-level observation data. This includes a list of visible satellites with information on their line-of-sight direction and received signal strength (\$GPGSV) or the range residuals of all satellites used in the navigation solution (\$GPGRS). While originally conceived for GPS receivers, the current NMEA 0183 standard also considers other GNSS constellations (GLONASS, Galileo) through dedicated message types and satellite/signal identifiers.

While substantial documentation on the format of the most popular NMEA 0183 GPS messages is available from public Internet sources, users should consider the official standard [A.3] as their primary source of information. It can be obtained directly from the NMEA web site [A.2] at a fee contributed to cover the work of this organization.

-----1 0-----2 0-----3 0-----4 0-----5 0-----6 0-----7 0-----8															
3.03		OBSERVATION DATA					M		RINEX VERSION / TYPE						
sbf2rin-8.5.1		ESA/ESOC					20131107 000707 LCL		PGM / RUN BY / DATE						
kour									MARKER NAME						
97301M210									MARKER NUMBER						
Automatic		ESA/ESOC							OBSERVER / AGENCY						
3001301		SEPT POLARX4					2.5.2		REC # / TYPE / VERS						
5129		SEPCHOKE_MC					NONE		ANT # / TYPE						
3839591.4332		-5059567.5514					579956.9164		APPROX POSITION XYZ						
0.0950		0.0000					0.0000		ANTENNA: DELTA H/E/N						
G	18	C1C	L1C	D1C	S1C	C1W	S1W	C2W	L2W	D2W	S2W	C2L	L2L	D2L	SYS / # / OBS TYPES
		S2L	C5Q	L5Q	D5Q	S5Q									SYS / # / OBS TYPES
E	16	C1C	L1C	D1C	S1C	C5Q	L5Q	D5Q	S5Q	C7Q	L7Q	D7Q	S7Q	C8Q	SYS / # / OBS TYPES
		L8Q	D8Q	S8Q											SYS / # / OBS TYPES
R	12	C1C	L1C	D1C	S1C	C2P	L2P	D2P	S2P	C2C	L2C	D2C	S2C		SYS / # / OBS TYPES
C	8	C2I	L2I	D2I	S2I	C7I	L7I	D7I	S7I						SYS / # / OBS TYPES

30.000									INTERVAL						Omitted lines
2013	11	6	0	0	0.0000000	GPS	TIME OF FIRST OBS								
2013	11	6	23	59	30.0000000	GPS	TIME OF LAST OBS								

															Omitted lines
															END OF HEADER
-----1 0-----2 0-----3 0-----4 0-----5 0-----6 0-----7 0-----8															

Start of header

Site meta data

Observation types

Epoch range

End of header

Fig. A.2 Sample RINEX 3 observation file header from the IGS KOUR station in Kourou, French Guyana. Rulers in the first and last line have been added for illustration only and are not part of the actual file. Colors indicate different groups of header records. Selected records from the original RINEX file have been omitted for brevity

Table A.3 RINEX 3 satellite system identifiers

ID	System
G	GPS
R	GLONASS
S	SBAS payload
E	Galileo
C	BeiDou
J	QZSS
I	IRNSS/NavIC

dex of the observation types provided later for the satellites of each GNSS constellation. The individual constellations are identified by a single-character identifier in column 1 of these records. Currently supported systems and designations are summarized in Table A.3. Except for GPS (G) and SBAS (S) the satellite system identifiers are derived from the nation (Russia, Europe, China, Japan, India) operating the specific constellation.

The individual observations types for each constellation are described by a three-character observation code. Its first character identifies one of the four types of measurements defined in Table A.4. Even though additional pseudo-observations such as channel number (X) and ionospheric phase delay (I) are formally permitted by the RINEX 3 standard, they are rarely used by the community.

Table A.4 RINEX 3 observation types and units

ID	Observation	Units
C	Pseudorange	m
L	Carrier-phase	cy
D	Doppler	Hz
S	Signal strength (C/N_0)	dB-Hz

The second and third characters of the observation code indicate the frequency band (single-digit band number) and the specific modulation or tracking mode (single-letter attribute). Since many of the new and modernized signals comprise multiple components (e.g., an in-phase channel with navigation information and a dataless pilot code on the quadrature channel), distinct attributes are defined for the individual components (e.g., I, Q) and the combined tracking of both components (e.g., X). An overview of currently defined RINEX 3 observation codes is provided in Table A.5.

The observation codes listed in the SYS / # / OBS TYPES header records define the set of observations provided for satellites of a given constellation. In the Galileo example of Fig. A.2, a full set of four measurements (pseudorange, carrier-phase, Doppler, signal strength) is specified for tracking of the pilot components of the E1 Open Service (*1C), E5a (*5Q), E5b (*7Q) and E5 AltBOC (*8Q) signals.

Table A.5 RINEX 3 observation codes for carrier-phase observations. Corresponding observation codes with initial letters C, D, and S apply for pseudorange, Doppler and signal strength observations

System	Band	Code	Description
GPS	L1	L1C	C/A-code
		L1S, L1L, L1X	L1C (data, pilot, combined)
		L1P	P-code (unencrypted)
		L1W	Semicodeless P(Y) tracking
		L1Y	Y-code (with decryption)
		L1M	M-code
		L1N	codeless
	L2	L2C	C/A-code
		L2D	Semi-codeless P(Y) tracking (L1 C/A+(P2-P1))
		L2S, L2L, L2X	L2C-code (medium, long, combined)
		L2P	P-code (unencrypted)
		L2W	Semicodeless P(Y) tracking
		L2Y	Y-code (with decryption)
		L2M	M-code
		L2N	codeless
	L5	L5I, L5Q, L5X	L5 (data, pilot, combined)
GLONASS	L1	L1C	C/A-code
		L1P	P-code
	L2	L2C	C/A-code
		L2P	P-code
	L3	L3I, L3Q, L3X	L3 (data, pilot, combined)
SBAS	L1	L1C	C/A-code
	L5	L5I, L5Q, L5X	L5 (data, pilot, combined)
Galileo	E1	L1A	PRS signal
		L1B, L1C, L1X	OS (data, pilot, combined)
		L1Z	PRS + OS(data+pilot)
	E5a	L5I, L5Q, L5X	E5a (data, pilot, combined)
	E5b	L7I, L7Q, L7X	E5b (data, pilot, combined)
	E5	L8I, L8Q, L8X	E5 AltBOC (data, pilot, combined)
BeiDou (BDS-2)	B1	L2I, L2Q, L2X	B1I(OS), B1Q, combined
	B2	L7I, L7Q, L7X	B2I(OS), B2Q, combined
	B3	L6I, L6Q, L6X	B3I, B3Q, combined

Table A.5 (continued)

System	Band	Code	Description
QZSS	L1	L1C	C/A-code
		L1S, L1L, L1X	L1C (data, pilot, combined)
		L1Z	L1-SAIF signal
	L2	L2S, L2L, L2X	L2C-code (medium, long, combined)
IRNSS/NavIC	L5	L5I, L5Q, L5X	L5 (data, pilot, combined)
	E6	L6S, L6L, L6X	LEX signal (short, long, combined)
	L5	L5A	SPS Signal
		L5B, L5C, L5X	RS (data, pilot, combined)
	S	L9A	SPS signal
		L9B, L9C, L9X	RS (data, pilot, combined)

Later, in the observation section of the file, a total of 16 matching measurements will be provided for each Galileo satellite.

A (truncated) example of an observation record corresponding to the header of Fig. A.2 is shown in Fig. A.3. The record starts with an epoch line (marked by a “>” character in column 1) specifying the date and time of the observations (here: 6 Nov 2013 00:00:00.0 GPS time) and the number of tracked satellites (here: 20). Subsequently the observations are provided in a single line for each satellite. The individual satellites are identified in columns 1–3 by the satellite number. This is made up of the satellite system indicator and a two-digit number identifying the transmitted pseudorange random noise (PRN) code or, in case of GLONASS, the slot number of the tracked satellite. Each observation is stored in a fixed field of 14 characters and with three trailing digits after the decimal point. Carrier-phase and pseudorange observations are furthermore complemented by an optional loss-of-lock indicator (0, blank, or 1) and/or a single-digit signal-strength indicator in the two cells adjacent to the actual measurement value.

Although the RINEX format supports the loss-of-lock indicator for each observation type, it is common practice to only indicate it on the phase observations. Likewise, the single-digit signal-strength field is often omitted if the signal strength (in dB-Hz) is explicitly provided as an observation type.

Hatanaka Compression. A practical problem associated with the transfer and storage of RINEX observation data is the large file size caused by the use of

Epoch		No. of satellites		Truncated	
Satellite number		Signal strength and loss-of-lock indicators			
> 2013 11 06 00 00 0.0000000 0 20					
G11	21789794.010 7 114506100.36007	1720.652 7	47.500	21789793.282 5	35.000
G15	23437270.527 7 123163643.91607	-899.995 7	44.500	23437270.523 5	30.500
G25	23279919.760 6 122336912.64006	2828.062 6	38.250	23279919.300 4	25.750
G05	23600086.200 7 124019495.78607	-1485.409 7	43.500	23600086.267 4	29.750
R00	23006036.629 6 123023649.96106	-1272.486 6	36.000	23006052.039 6	95685099.64406
G24	21004445.599 8 110379323.36308	-1995.581 8	50.000	21004445.701 7	43.000
G18	23632299.040 7 124188483.90207	2852.612 7	43.750	23632298.312 4	25.750
G14	25391662.793 5 133434050.21305	442.853 5	35.000	25391662.346 2	12.000
R16	19300482.290 8 10309905.21108	145.853 8	48.250	19300491.076 7	80188907.52507
G12	21917255.393 8 115176179.65808	1821.021 8	50.000	21917255.309 6	38.250
G29	23768181.065 6 124902773.05306	-802.817 6	41.500	23768180.905 3	22.000
R09	21623927.772 7 115470629.18307	3608.897 7	46.750	21623935.553 6	89810480.72806
R06	24236049.702 7 129328208.84807	2479.388 7	44.250		
C14	24458718.261 7 127362937.75107	2345.948 7	44.250	24458727.629 8	98485138.61708
R15	21481799.314 7 114792351.97907	-3609.734 7	45.000	21481813.243 7	89282811.39807
R04	22254258.875 7 119170642.67807	174.797 7	44.750	22254267.166 6	92688257.97106
R19	23228568.363 7 124257247.38907	-2230.894 7	42.250	23228573.599 6	96644606.11206
R05	21839218.040 7 116743131.47507	2026.164 7	46.750	21839233.552 7	90800281.60607
C12	25394611.396 6 132236540.71606	-985.074 6	40.500	25394615.635 7	102253547.41407
C11	25458822.666 7 132570867.23307	-836.718 7	43.250	25458828.466 7	102512178.67707

Fig. A.3 Sample observation record from a RINEX observation file obtained by the IGS KOUR station in Kourou, French Guyana. Rulers in the first and last line have been added for illustration only and are not part of the actual file

an ASCII text format. Considering a 30 s sampling, a multi-GNSS reference station with 40 or more commonly tracked satellites provides daily RINEX 3 observations of about 25 MB and an even higher data volume is obtained with 1 Hz high-rate data. This problem can partly be alleviated through standard compression tools such as compress, zip, and rar, which achieve a compression ratio of 1 : 3 to 1 : 4 for typical RINEX files.

A complementary compression technique [A.9], which retains the ASCII representation but reduces the high level of redundancy in the observation data themselves has been developed by *Yuki Hatanaka* at the Geospatial Information Authority (GSI), Japan. It is based on the observation that a time series of consecutive measurements may better be represented by an initial value and the differences between epochs, since the differences are smaller in size than the absolute values and require a reduced number of digits.

Hatanaka compression retains most of the RINEX file header but replaces the observation records with the differenced data and uses a “&” separator between data fields to minimize the overall amount of white space. The resulting *Compact RINEX* format already achieves a file size reduction by a factor of 3–4. In combination with standard file compression tools, an overall reduction by a factor of approximately eight is achieved compared to the original RINEX file. It should be noted that the Hatanaka compression is lossless and enables a full recovery of the original observations data upon decompression.

Navigation Data. As a complement to the observation data format discussed above, the RINEX standard also defines a navigation data format for the exchange of broadcast ephemeris information. These broadcast ephemerides comprise the orbit and clock parameters as well as auxiliary data transmitted by each GNSS satellite for use in real-time navigation. Even though the broadcast ephemerides are generally less accurate than postprocessed ephemeris products, the navigation data files are frequently used for preprocessing of GNSS data (e.g., elevation screening), relative positioning, and as a priori information for precise GNSS orbit and clock determination.

While distinct GPS and GLONASS files were foreseen in early versions, the current RINEX 3 standard supports mixed files with navigation data for all GNSSs. Similar to the observation data format, navigation files are fixed-format text files but are limited to 80 characters per line. The file header follows the same concepts as introduced before and starts with two lines identifying the file type and format version. For the purpose of illustration an annotated format example is given in Fig. A.4.

All header parameters are optional and may comprise different types of ionospheric model parameters and time conversion parameters. These include the eight coefficients $\alpha_0, \dots, \alpha_3$ and β_0, \dots, β_3 of the Klobuchar model for GPS and QZSS users or the Klobuchar-style model employed by BeiDou. For Galileo, a set of four parameters A_0, \dots, A_3 is provided that enables ionospheric corrections of single-frequency observations with the NeQuick model. An overview of these models is provided in Chap. 6 and references therein, while detailed information on the application of the broadcast ionospheric parameters is given in the interface specifications of the individual GNSSs [A.10–14].

The remaining header parameters provide the relation of different GNSS timescales among each other and with UTC. These include the number of UTC leapseconds since 1980 (or, equivalently the integer seconds time difference between GPS time and UTC) as well as linear polynomials for the fractional time differences (which typically amount to some tens or at most hundreds of nanoseconds). The individual time correction parameters are identified by a 2 + 2 character code in columns 1–4 of the TIME SYSTEM CORR header lines and contain the polynomial coefficients (A_{f0}, A_{f1}) as well as the corresponding reference epoch (week and seconds of week). While not a mandatory header line, provision of the leapseconds information is strongly recommended since it facilitates the translation between the native time systems of each constellation and enables the processing of the subsequent ephemeris data in a consistent timescale.

Following the header, a series of ephemeris data sets for individual satellites and epochs are given. Following the satellite identification and reference epoch, the various ephemeris parameters are provided in fixed fields of 19 characters. The set of parameters and their total is predefined for each individual constellation (GPS, GLONASS, BeiDou, Galileo, QZSS, IRNSS/NavIC, and SBAS) and reflects the different types of information made available in the various types of navigation messages.

Common to all constellations, the satellite clock offset is specified through a clock reference epoch (t_{oc}) and a clock offset polynomial (a_{f0}, a_{f1} , and, for most systems, a_{f2}). For many GNSSs, the clock offset information is complemented by differential code biases (known as timing group delay (TGD), broadcast group delay (BGD) or intersignal correction (ISC) parameters). These enable a consistent processing of the transmitted clock offsets when using single-frequency observations or a different set of signals than the ones used by the control segment to determine the broadcast clock offsets. Clock-related parameters in the format example of Fig. A.4 are highlighted in blue color.

```

-----1|0-----2|0-----3|0-----4|0-----5|0-----6|0-----7|0-----8|
3.03      NAVIGATION DATA      M (Mixed)      RINEX VERSION / TYPE
BCEmerge  MGEX                  20131107 044603 GMT PGM / RUN BY / DATE
GPSA      0.2515e-07 -0.7451e-08 -0.1192e-06  0.1192e-06  IONOSPHERIC CORR
GPSB      0.1331e+06 -0.4915e+05 -0.1311e+06 -0.1311e+06  IONOSPHERIC CORR
GAL       1.2525e+02 -5.7031e-01  1.0834e-02  0.0000e+00  IONOSPHERIC CORR
QZSA      0.5122e-07 -0.3353e-06 -0.1192e-06  0.3517e-05  IONOSPHERIC CORR
QZSB      0.1536e+06 -0.8356e+06  0.4063e+07 -0.6554e+07  IONOSPHERIC CORR
GAUT      1.3969838619e-08-5.329070518e-15 172800 1765      0 TIME SYSTEM CORR
GLGP      -3.8184225559e-07 0.000000000e+00 259200 1765      0 TIME SYSTEM CORR
GLUT      -1.8579885364e-07 0.000000000e+00 259200 1765      0 TIME SYSTEM CORR
GPGA      9.6333678812e-09 5.329070518e-15 345600 1757      0 TIME SYSTEM CORR
GPUT      -1.8626451492e-08-0.532907052e-14 405504 1765      0 TIME SYSTEM CORR
QZUT      2.7939677238e-08-3.552713680e-14 491520 1765      0 TIME SYSTEM CORR
16      16 1694      7
LEAP SECONDS
END OF HEADER
-----Omitted lines-----
E12 2013 11 06 02 20 00 3.416970139369e-05 1.233502189280e-11-1.734723475977e-18
6.200000000000e+01-1.281250000000e+00 3.154417108420e-09-1.584524137928e+00
-2.607703208923e-08 9.329034946859e-05 1.055561006069e-05 5.440608764648e+03
2.676000000000e+05 5.774199962616e-08 1.159805493823e+00 5.587935447693e-09
9.589155555733e-01 1.139062500000e+02-2.779965207269e+00-5.471656487884e-09
7.568172387615e-10 5.130000000000e+02 1.765000000000e+03
-1.000000000000e+00 3.900000000000e+02-2.793967723846e-09-2.561137080193e-09
2.682550000000e+05
-----Omitted lines-----
R03 2013 11 06 02 15 00 2.223066985607e-06 0.000000000000e+00 2.664000000000e+05
-1.547962890625e+03-2.572350502014e+00-3.725290298462e-09 0.000000000000e+00
1.293956201172e+04 1.558908462524e+00-0.000000000000e+00 5.000000000000e+00
-2.189239355469e+04 1.099916458130e+00 9.313225746155e-10 0.000000000000e+00
-----Omitted lines-----
-----1|0-----2|0-----3|0-----4|0-----5|0-----6|0-----7|0-----8|

```

Start of header

Ionosphere
parametersTime system
parameters

End of header

Galileo
ephemerisGLONASS
ephemeris

Field	1	2	3	4	5	Row
Sat	Epoch (t_{oc} , GPS time)	a_{f0} (s)	a_{f1} (s/s)	a_{f2} (s/s ²)		0
	IOD _{nav}	C_{rs} (m)	Δn (rad/s)	M_0 (rad)		1
	C_{uc} (rad)	e	C_{us} (rad)	\sqrt{a} (m ^{1/2})		2
	t_{oc} (s)	C_{ic} (rad)	Ω_0 (rad)	C_{is} (rad)		3
	i_0 (rad)	C_{rs} (m)	ω (rad)	$d\Omega/dt$ (rad/s)		4
	di/dt (rad/s)	Data source	Week			5
	Accuracy (m)	Health	BGD _{E5aE1} (sec)	BGD _{E5bE1} (sec)		6
	Transmission time (s)					7

Field	1	2	3	4	5	Row
Sat	Epoch (t_{oc} , UTC)	$a_{f0} = -\tau_N$ (s)	$a_{f1} = +\Gamma_N$ (s/s)	Message frame time (s)		0
	x (km)	dx/dt (km/s)	d^2x/dt^2 (km/s ²)	Health		1
	y (km)	dy/dt (km/s)	d^2y/dt^2 (km/s ²)	Frequency channel k		2
	z (km)	dz/dt (km/s)	d^2z/dt^2 (km/s ²)	Age of information		3

Fig. A.4 Format example of a multi-GNSS RINEX navigation file (*top*). Layout of Galileo and GLONASS ephemeris records (*bottom*; clock and orbit parameters are highlighted by blue and red color)

Orbit information parameters provided in the individual broadcast navigation data vary among the different constellations, but fall in either of two basic

categories: orbital elements (\sqrt{a} , e , i_0 , Ω_0 , ω_0 , M_0) and perturbation coefficients (di/dt , $d\Omega/dt$, C_{rc} , C_{rs} , C_{ic} , C_{uc} , C_{us}) for use with a perturbed Keplerian orbit

0xD3	Res.	n	Message [0, ..., $n-1$]		CRC-24
			Msg.No.	Other data	

Fig. A.5 RTCM SC-104 v3.x message protocol. The first field contains the 8 bit preamble (hex value 0xD3)

model or Cartesian state vectors $(x, y, z, \dot{x}, \dot{y}, \dot{z}, \ddot{x}, \ddot{y}, \ddot{z})$ for numerical integration or polynomial interpolation of the trajectory across short time intervals. Details of the orbit models and the employed parameters are provided in Chap. 3 as well as the interface specifications of the individual constellations [A.10–15]. Broadcast ephemerides with orbital elements are presently used for GPS, QZSS, Galileo, BeiDou, and IRNSS/NavIC while state vectors are provided by the GLONASS and SBAS satellites. The corresponding parameters are highlighted using red color in the format example of Fig. A.4.

Up to version 3.03 the RINEX navigation format supports only legacy navigation data (i.e., one parameter set per constellation), but will be extended for CNAV, CNAV2 and other modernized broadcast ephemeris data sets in subsequent releases.

A.1.3 RTCM SC-104 DGNSS Data Format

The Radio Technical Commission for Maritime Services (RTCM) was formed as a United States government advisory committee in 1947. Currently, the RTCM is an international nonprofit scientific, professional and educational organization that is supported by its members from all over the world. Membership consists of both corporate and government organizations. RTCM members charter Special Committees (SC) to address in-depth radio communication and radio navigation issues, with the goal of supporting interoperability.

RTCM Special Committee 104 (RTCM-SC104) was chartered to address differential global navigation satellite systems (DGNSSs). Initially, RTCM SC-104 focused on standards and protocols for differential GPS for maritime applications. SC-104's mandate has grown to support not only maritime differential GNSS, but also real-time kinematic and precise GNSS data formats and Network Transport of RTCM using Internet Protocol (NTRIP, [A.16]).

The following paragraphs provide an overview of the RTCM SC-104 standard for DGNSS services based on the latest version 3.3 [A.17]. A full specification is available from the RTCM [A.18] at a service fee, which contributes to covering the work of this organization.

Message Types and Format. RTCM SC-104 messages are primarily designed for real-time GNSS applications and use a binary protocol to minimize the overall data volume that needs to be transferred between providers and users. While early versions made

use of a message format made up of fixed-length data words with parity protection similar to that of the GPS navigation message, a revised, variable-length message format was introduced for use from version 3.0 onwards. It comprises a header with an 8 bit preamble, 6 bit reserved fields and a 10 bit message length field (Fig. A.5).

Following the header, a data field of up to 1023 bit is provided and the message concludes with a 24 bit cyclic redundancy check (CRC) checksum to ensure integrity of the transmitted data. Except for zero-length filler messages, all messages start with a 12 bit message number that identifies the message type and provides the key to decoding the subsequent message data fields.

RTCM SC-104 v3.x defines various groups of messages for observation data, network RTK corrections, auxiliary and metadata as well as state space correction data (Table A.6). Various multisignal messages provide similar parameters albeit in different combinations and/or resolutions. In this way positioning services of different accuracy may be implemented that make best use of the available communication bandwidth.

Multisignal Messages. Early versions of the RTCM SC-104 standard were strongly focused on the GPS and GLONASS systems and their legacy L1/L2 signals. As of version 3.2, the concept of multisignal messages (MSMs, [A.19]) has been introduced, to establish a truly generic framework for observations of all GNSS constellations and all transmitted signals.

The multisignal messages employ a highly efficient packing scheme, which minimizes the overall amount of data that need to be transmitted. For each observed satellite, the pseudorange and carrier-phase observations are decomposed into the sum of a *rough range* and a *fine range*. The rough range at each epoch is common to all observations of the given satellite collected simultaneously on the individual signals. The remaining fine range reflects the differences in ionospheric path delays and systematic GNSS satellite and receiver biases. It is generally confined to less than ± 300 m, which requires considerably fewer data bits for storage than the original value.

Other MSM features include the signal and satellite masks, which serve as an index for the overall set of tracked signals (across all satellites of a constellation) and the subset tracked for a specific satellite. In this way, the presence of modernized signals available for only part of a constellation (e.g., L2C on GPS Block

Table A.6 RTCM SC-104 v3 message groups

Groups	Type	Messages
Experimental messages		0–100
Observations	GPS L1, L1/L2	1001–1004
	GLONASS L1, L1/L2	1009–1012
	Multi Signal Messages for individual GNSSs	1071–1077 (GPS) 1081–1087 (GLONASS) 1091–1097 (Galileo) 1101–1107 (SBAS) 1111–1117 (QZSS) 1121–1127 (BeiDou)
Site metadata	Station coordinates, receiver and antenna description	1005–1008, 1032–1033
Network RTK corrections	Auxiliary station data	1014
	Geometric and ionospheric corrections	1015–1017, 1037–1039
	Network RTK residuals and FKP gradients	1030–1031, 1034–1035
Auxiliary information	System parameters	1013
	Satellite ephemeris	1019, 1020, 1042, 1044, 1045, 1046
	Text (unicode)	1029
	GLONASS code/phase biases	1230
Transformation parameters		1021–1027
State Space Representation parameters	Orbit & clock corrections, code biases, URA	1057–1062 (GPS) 1063–1068 (GLONASS)
Proprietary messages		4001–4095

IIR-M and L5 on Block IIF) can be supported without requiring empty or padded data fields for older satellites.

Finally, a total of seven different multisignal messages are defined for each constellation. The individual message types are distinguished by the final digit of the message number and offer different sets of elementary observations (pseudorange, carrier-phase, signal strength, and Doppler) at different levels (compact, full) of range and resolution.

Network-RTK Messages. Network-RTK refers to real-time kinematic positioning (RTK) using correction data derived from a network of terrestrial reference stations (Chaps. 26 and 31). Compared to a single reference station, the network-based corrections can be applied in a wider region and the quality of carrier-phase ambiguity resolution becomes less dependent on the base station distance. The RTCM SC-104 standard offers a harmonized framework for transmitting such corrections to the user independent of the underlying network architecture. Both GPS and GLONASS network-RTK services are supported through dedicated messages. Aside from station-related information provided in the auxiliary station data message, *geometric* (nondispersive) and *ionospheric* (dispersive) correction data may be transmitted in distinct or combined messages. Details of the respective messages and the appli-

cable processing conventions are provided in the RTCM SC-104 standard [A.17]. Another class of network-RTK messages comprises the *Network-RTK Residual Error* messages, which are used to implement concepts such as virtual base stations to improve the RTK service for specific users.

State-Space Representation Messages. The state-space representation (SSR) represents a new concept for the provision of correction data in real-time kinematic precise point positioning (PPP-RTK) applications. Rather than providing combined corrections in observation space, the SSR approach employs decomposed corrections to remove individual GNSS error sources [A.20]. These include satellite position corrections (in three dimensions) and satellite clock corrections as well as code biases. Different message types are supported for individual or combined orbit and clock corrections. Furthermore, distinct high-rate clock correction messages are available to ensure that satellites with fast changing atomic clocks can be accurately characterized. The SSR concept also foresees the provision of vertical total electron content (VTEC) information for single-frequency users, even though these are not yet part of the RTCM SC-104 standard. The generic nature of the SSR corrections makes them largely independent from the user location and provides the basis for global PPP applications.

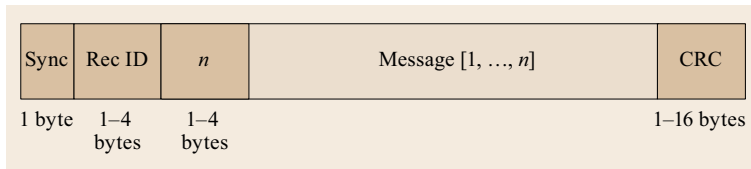


Fig. A.6 Basic BINEX message protocol

A.1.4 GNSS BINary Exchange (BINEX) Format

The BINary Exchange (BINEX) format, is a GNSS format standard that supports both research and operational applications. BINEX was developed at UNAVCO under the leadership of *Lou Estey* (UNAVCO) to achieve a better data compression and increased reliability for real-time data streams from permanent GPS monitoring stations [A.21, 22]. Other than common binary formats, BINEX is designed to support both little-endian and big-endian word-orders, which allows use on a wide range of hardware platforms and processors. Also, a wide range of message lengths is supported through checksum and message number fields with a varying number of bytes. The BINEX format supports observation and navigation messages for all current GNSS constellations as well as meta-data messages to encapsulate site-specific parameters. BINEX is supported by major GNSS receiver manufacturers (Trimble, Topcon, Javad, Leica, Septentrio) and is continuously extended and refined to meet the needs of new signals and systems. The official documentation of the BINEX format is maintained as a living document on the UNAVCO web site [A.23].

Record Structure. BINEX data files are made up of a sequence of consecutive BINEX data records that can be processed independent of each other. Since BINEX files do not contain a file header, multiple BINEX files can easily be concatenated for further processing without the need for special tools. The BINEX standard supports various generic forms of message frames for different applications. These are designed to enable any foreseeable type of GNSS receiver data, auxiliary information as well as final data products.

In their most simple (and widely used) form, each record comprises a header with synchronization byte, record identifier and message length field before the actual message data and concludes with a checksum field (Fig. A.6). The message itself may contain additional fields (subrecord ID, etc.) to further distinguish the actual contents.

Both the record identifier and the message length are stored in an *unsigned BINEX integer* (ubnxi) data word, which occupies between 1–4 byte depending on the size of the encoded value. Likewise, different types of checksum (1 byte XOR checksum, 2 byte CRC-16, or

Table A.7 Common BINEX record types

ID	Sub ID	Contents
0x00		Site Metadata
0x01		GNSS Navigation Information
	0x00	● coded (raw bytes) GNSS ephemeris
	0x01	● decoded GPS ephemeris
	0x02	● decoded GLONASS-FDMA ephemeris
	0x03	● decoded SBAS ephemeris
	0x04	● decoded Galileo ephemeris
	0x05	● decoded Beidou-2/Compass ephemeris
	0x06	● decoded QZSS ephemeris
	0x07	● decoded IRNSS ephemeris
	0x41–0x47	● raw navigation subframe/block/page for individual constellations
0x7d		Receiver Internal State
	0x00	● Temperature and power
0x7e		Ancillary Site Data Prototyping
	0x00	● Meteorological and local geophysical data
0x7f		GNSS Observable Prototyping
	0x00	● for JPL LEO support network
	0x01	● for UCAR COSMIC and GPS/MET
	0x02	● for UCAR Suominet
	0x03	● for EarthScope
	0x04	● for EarthScope
	0x05	● Generic multi-GNSS observation data

4 byte CRC-32) are employed depending on the length of the message data field.

Message Types. While BINEX is a highly generic protocol intended to support the binary transfer of all types of GNSS-related information, only a limited number of messages are widely used at present. All of these employ a 1 byte record identifier and belong to one of five major categories shown in Table A.7.

For most record types, the message fields start with a 1 byte subrecord ID to further distinguish the data contained within. Even though records 0x7e and 0x7f are formally considered as prototype messages that shall eventually be replaced by 0x03 and 0x02 messages, some of them (such as 0x7f 0x05) represent a de facto standard and have already been adopted by various receiver manufacturers.

In addition to the public messages described in Table A.7, various record identifiers have been assigned to different institutions or companies to enable the implementation of private BINEX messages. These make use of record IDs beyond 0x7f (= 127) that require more than one byte for ubnxi encoding.

A.2 IGS Product and Metadata Formats

Complementary to the text and binary formats for exchange of receiver-related information (observations, navigation data, metadata, etc.) a wide range of different formats have been developed in the frame of the International GNSS Service (IGS). These enable a consistent exchange of GNSS data products and auxiliary information among users and analysis centers. Most common examples include precise orbit and clock information, atmospheric products, antenna information and site metadata, all of which are described in this section. Current and past versions of the various format specifications are available in electronic form through the IGS web site [A.8].

A.2.1 SP3 Ephemeris Format

The *Standard Product 3 (SP3)* format defines a widely used standard for the provision of precise orbit and clock data of GNSS satellites. Aside from RINEX observation data, the SP3 orbit and clock information forms the basis of most precise point positioning (PPP) applications. IGS orbit and clock products are consistently made available in SP3 format by the various analysis centers and can be utilized by all common PPP software packages.

SP3 originates from two types of GPS orbit data formats (SP1, SP2) developed by the US National Geodetic Survey (NGS). Other than its predecessors that were limited to orbit-only data, SP3 also incorporates clock data and accuracy information [A.24]. Even though both text and binary versions have originally been defined, only the former has found widespread acceptance and continues to be developed. In its latest version, SP3d, the format supports all GNSS constellations using satellite identifiers consistent with those of the RINEX standard (Sect. A.1.2). To facilitate a joint use for precise orbit information of satellites in low Earth orbit (LEO), a constellation letter “L” has furthermore been introduced for non-GNSS satellites in addition to the designations in Table A.3.

Even though the SP3 format is specifically designed to combine orbit and clock offset information in a fully consistent manner, clock data are often desired at higher rate. In view of their smooth motion, GNSS orbits with periods of 12–24 h can readily be interpolated from known values at a 15 min spacing

or more. Clock variations, on the other hand, are governed by stochastic processes and a smaller sampling interval (down to 30 s or less) is required for accurate interpolation. PPP users may therefore prefer to complement SP3 orbit data with separate high-rate clock data (Sect. A.2.2), provided that both products have been generated by a common provider, using fully consistent processes.

Format Description. The basic format and contents of an SP3d orbit and clock data file are illustrated in Fig. A.7 based on the comprehensive specification in [A.25]. It comprises a header section with auxiliary information for the proper processing of the subsequent data records. These provide orbit and clock data as well as optional accuracy and event information on an equidistant epoch grid for a previously specified number of satellites and epochs. While the format was originally limited to a line width of 80 characters and a 22-line header, a larger number of header lines has been introduced in SP3d to accommodate more than 85 satellites and extended comments.

The file header comprises various blocks of lines (introduced by #, +, %, and / characters), which provide relevant indices and parameters for the proper interpretation of the subsequent orbit, clock and accuracy data records.

The first header line indicates the format version (“d” for SP3d in column 2) and distinguishes position-only files from files with position and velocity information (“P” or “V” in column 3). Thereafter the calendar date of the initial epoch and the number of data points is provided. Along with the stepsize provided in columns 25–38 of line 2, the epochs of all subsequent orbit and clock data records are fully defined by this header information. Complementary to the specification of the start epoch in line 1, a (redundant) representation in terms of weeks and seconds as well as integer and fractional Modified Julian Day count are given in line 2. Further information in the first header includes an indicator for the employed data (e.g., u+U for undifferenced carrier-phase and code observations), the coordinate system descriptor, the type of orbit product and an acronym of the responsible institution.

Following these initial header lines, a block of five or more lines introduced by a single + character specifies the total number of satellites, the satellite identifiers (constellation letter plus number) and the sequence of satellites for which orbit and clock data are later given in the epochwise SP3 data blocks. The next header block (highlighted in red in Fig. A.7) provides integer-valued orbit accuracy indicators a for each spacecraft, from which the standard deviation $\sigma_{\text{orb}} = 2^a$ mm of the respective orbit errors across all epochs can be obtained.

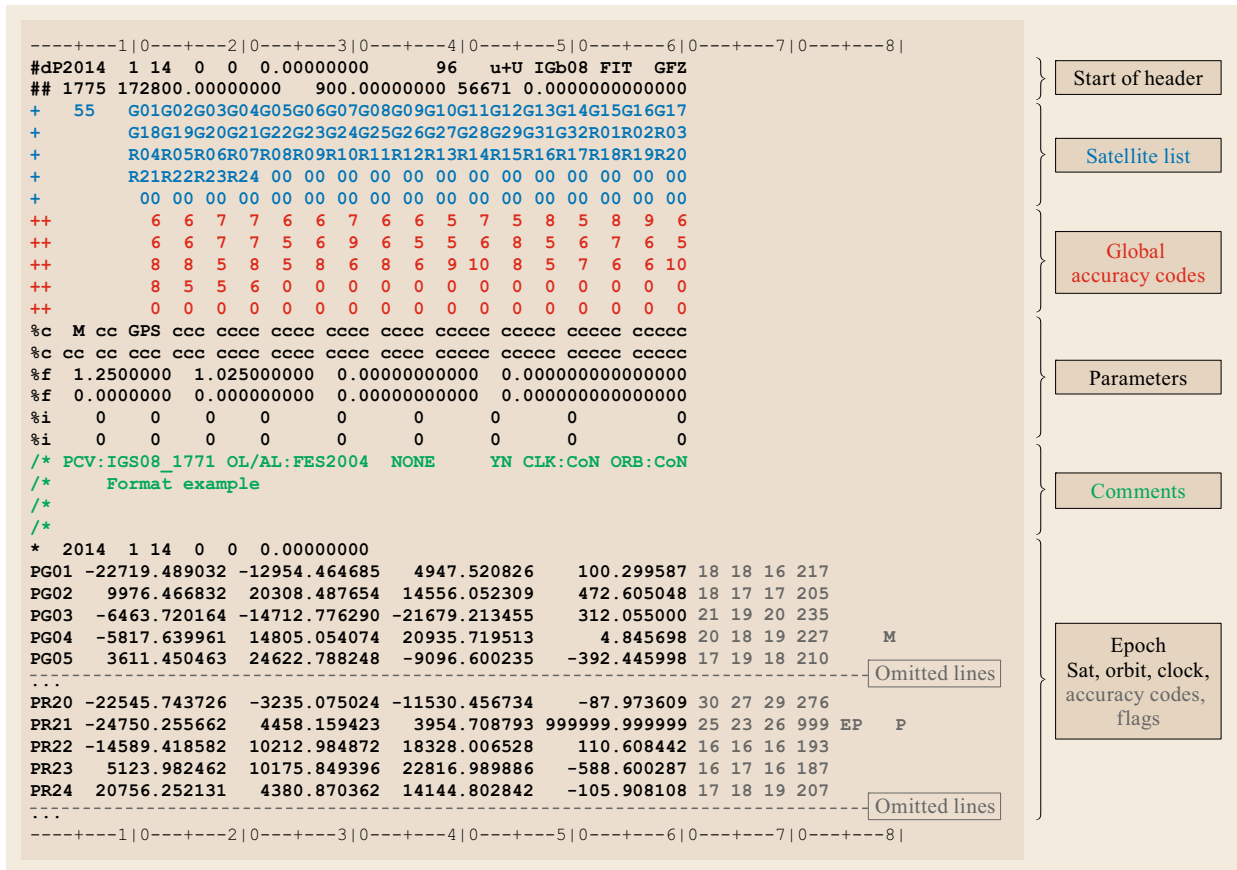


Fig. A.7 SP3d format example

The header continues with a block of auxiliary parameters, most notably the time system indicator in the first line starting with %c. It concludes with a block of four or more comment lines used within the IGS to identify various processing conventions such as antenna offset or tide models.

Each of the subsequent data blocks starts with an epoch header line specifying the epoch date and time, which must be consistent with the sequence of the record and the epoch grid defined in the file header. Thereafter, the position (in units of km) and clock offset (in μs) are specified for each satellite in lines marked by an initial “P” character. For files containing both position and velocity information, each position record is complemented by a subsequent velocity record (indicated by the initial “V” character), which provides the spacecraft velocity (in units of 0.1 m/s) and clock rate (in $10^{-4} \mu\text{s}$). Unknown or invalid data in both records are indicated by zero values (position and velocity) or 999999.999999 (clock offset and clock rate).

Fig. A.7 also illustrates the use of optional epoch- and componentwise accuracy indicators (columns

62–73) as well as additional flags (columns 75–80 for clock events (“E”), orbit maneuvers (“M”) and predicted orbit or clock data (“P”). Similar to the global accuracy indicators described above, these indicators provide the exponent a for computing the standard deviation of the respective value using a relation of the form $\sigma = b^a$. Other than using a fixed base of $b = 2$, the base value can be selected by the provider and is separately specified in the %F header line for position and clock data. In this way a better resolution of the accuracy information is achieved. Using epochwise accuracy indicators is optional but enables a better distinction of predicted orbit and clock information from values based on actual observations.

For completeness, we note that the SP3 standard also foresees the use of an optional position and clock correlation record (EP) as well as a corresponding EV record for velocity and clock rate information. These records enable provision of a fully four-dimensional covariance matrix to describe the statistical properties of the respective data. However, neither the position nor the clock correlation records

have found widespread acceptance in official IGS products.

Interpolation. Based on its history, SP3 is designed as a tabular ephemeris format, which provides all data at equal intervals. This is a natural choice for orbit information generated by numerical orbit prediction software and facilitates the interpolation of intermediate values. Having said that, some synchronization may be required to ensure that clock data can be made available at the same epochs as the orbit information. Even though the SP3 format supports the joint provision of epochwise position and velocity data, it has been recognized that position-only information is generally sufficient, since velocity can be obtained from these data through differentiation of an interpolating function. Versions with and without complementary velocity (and clock rate) data are therefore supported by the SP3 format. Aside from their reduced size, position-only SP3 files ensure full consistency of the derived position-velocity information and avoid the cumbersome transformation of inertial to Earth-fixed velocity data.

Various forms of interpolators for GNSS ephemeris data have been proposed and studied in the literature [A.26–28], but polynomial interpolation is probably most widely used. A variety of algorithms have been developed for this purpose [A.29], among which the Lagrange method appears best suited when multiple values (such as the x -, y -, and z -coordinates of the position vector) have to be interpolated on the same epoch grid. It is therefore commonly recommended for use with SP3 orbit data [A.24, 30, 31].

Given a set of $n + 1$ epochs t_i , the elementary n th-order Lagrange polynomials

$$l_i(t) = \prod_{j=0, j \neq i}^n \frac{(t - t_j)}{(t_i - t_j)} \quad (i = 0, \dots, n) \quad (\text{A.1})$$

are first computed. These are designed to vanish at all but one grid point, i. e.,

$$l_i(t_j) = \begin{cases} 1 \\ 0 \end{cases} \quad \text{for} \quad \begin{cases} i = j \\ i \neq j \end{cases}. \quad (\text{A.2})$$

With this result the interpolating polynomial of order n can conveniently be expressed as a linear combination

$$\mathbf{r}(t) = \sum_{i=0}^n \mathbf{r}_i l_i(t), \quad (\text{A.3})$$

where \mathbf{r}_i denotes the values of the position vectors at the given grid points. The velocity at time t can likewise be

obtained through Lagrange interpolation

$$\mathbf{v}(t) = \sum_{i=0}^n \mathbf{v}_i l_i(t), \quad (\text{A.4})$$

based on known values \mathbf{v}_i at the grid epochs. Alternatively, the interpolating polynomial (A.3) may be differenced to obtain the relation

$$\mathbf{v}(t) = \sum_{i=0}^n \mathbf{r}_i l'_i(t), \quad (\text{A.5})$$

where

$$l'_i(t) = \sum_{\substack{k=0 \\ k \neq i}}^n \frac{1}{(t - t_k)} \cdot \prod_{\substack{j=0 \\ j \neq i}}^n \frac{(t - t_j)}{(t_i - t_j)} \quad (\text{A.6})$$

denotes the time derivative of the n th-order Lagrange polynomial. It may be noted that the above expressions are also applicable for interpolation of nonequidistant values. However, a constant step size (as implied by the SP3 format) simplifies the computation of the Lagrange polynomials and contributes to an even distribution of interpolation errors.

At the 15 min spacing adopted for most GNSS orbit products, a ninth-order polynomial using an equal number of grid points on both sides of the interpolation epoch offers an interpolation accuracy compatible with the numerical resolution of the SP3 orbit data [A.26]. However, larger interpolation errors may be encountered near the beginning or end of the ephemeris period [A.27]. Also, high-order interpolation must not be applied to clock data, which are not as smooth as the orbit information. Here, linear interpolation or at most cubic interpolation [A.32] is recommended.

Conventions. To enable a consistent use of multi-GNSS orbit and clock files, the SP3 standard requires that information for all satellites shall refer to a common time and reference system identified in the file header. Despite this, further specifications by the product provider will typically be required for the proper interpretation of the data.

Within the IGS, GNSS satellite positions provided in the SP3 precise ephemeris products are referred to the spacecraft center-of-mass and a mean-crust-fixed terrestrial coordinate system (which may differ by a few millimeters from a barycentric system due to tidal deformation of the Earth).

Clock data, in contrast, refer to an adopted antenna phase center and a conventional dual-frequency signal combination (L1/L2 P(Y)-code observations for GPS).

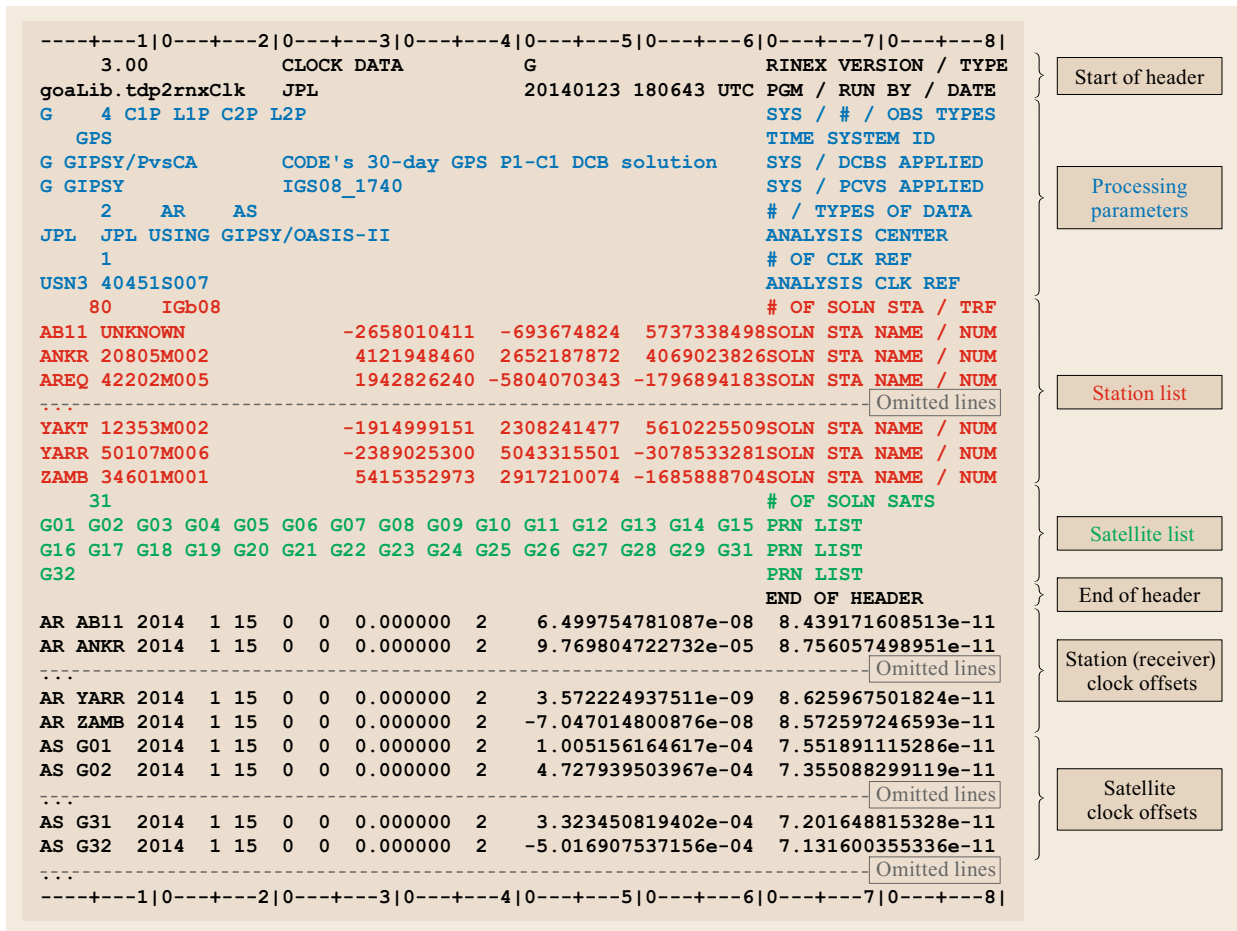


Fig. A.8 RINEX clock format example

Furthermore, the dominant relativistic contribution to the apparent clock caused by the orbital eccentricity has been removed from the provided clock data to better reflect the clock’s proper time. As such, a corresponding correction must be added to clock offsets retrieved from the SP3 ephemeris for the proper modeling of GNSS observations (Sect. 19.2).

A.2.2 Clock RINEX Format

The *RINEX Extensions to Handle Clock Information* [A.33] were introduced in 1998 as a generic framework for the exchange of general clock offset information. Today, the format is primarily used to provide precise GNSS satellite and receiver clock data derived from the analysis of global networks of monitoring stations. Despite its name, the Clock RINEX format has therefore evolved to a GNSS product format rather than a receiver data format. Clock RINEX data are widely used for precision point positioning (PPP) and also to monitor the performance of GNSS satellite clocks

and atomic clocks of national time standard laboratories.

The structure and contents of a Clock RINEX file are illustrated in Fig. A.8. In accordance with RINEX observation and navigation data formats (Sect. A.1.2), the clock data file header is made up of a sequence of header lines identified by their labels in columns 61–80. These mandatory or optional header lines provides auxiliary information for the proper interpretation of the actual clock data and are terminated by a `END OF HEADER` line. Key parameters provided within the header comprise a time system indicator, information on the signal or signal combination used in the clock offset determination for each GNSS constellation, information on the application of phase center offsets, phase pattern variations and differential code biases, and the station clock serving as a reference for all clock offsets. Additionally, lists of all stations and satellites used in the clock offset estimation process are provided.

The subsequent data section of the RINEX clock file provides clock offset values for each of these stations and receivers on an epoch-by-epoch basis. The respective lines are marked by a leading “AR” (for analysis data, receivers) and “AS” (for analysis data, satellites) and contain explicit epoch information. The remaining fields specify the number of data items followed by the clock offset and, optionally, its standard deviation. Both values are provided in units of seconds. If desired, clock rate and acceleration as well as their standard deviation may also be given on a continuation line.

A fixed stepsize is not required but it is a common practice among the IGS analysis centers. Depending on the product and provider, step sizes of 5 min and 30 s are most frequently used, but 5 s products are also available for high-precision applications. Linear interpolation of clock data between consecutive epochs is recommended in view of the stochastic nature of typical clock variations.

A.2.3 SINEX

The Solution INdependent EXchange (SINEX) format [A.34] was originally designed to facilitate the exchange and distribution of station coordinates, velocities and Earth orientation parameters (EOP) between the IGS analysis centers. These parameters are routinely estimated from the processing of the global IGS GNSS network using a wide range of software tools. A common format was therefore required to enable a comparison and combination of the analysis center solutions. SINEX has been used for this purpose from 1995 onwards and received continuous amendments to handle new parameters over the past decades.

The format has also been adopted and extended to meet the needs of Very Long Baseline Interferometry (VLBI), Satellite Laser Ranging (SLR) and Doppler Orbitography and Radiopositioning Integrated by Satellite (DORIS) techniques. The combination of these techniques is used to generate the International Terrestrial Reference Frame (ITRF).

A basic SINEX file contains station coordinate and Earth rotation parameter estimates along with auxiliary information about the site, such as receiver and antenna type, eccentricity and phase center information. In addition, covariance information or normal equations can be stored to facilitate the combination of solutions from different analysis centers and/or geodetic observation systems [A.35].

SINEX employs a fixed-format text representation with a maximum line width of 80 characters. Each file is started with a mandatory `%=SNX` header line and terminated with an `%ENDSNX` footer. In between, various blocks of data may be given in arbitrary order. Each block is identified by a predefined label and embed-

ded in a frame of `+label ... -label` lines. An overview of SINEX data blocks file is given in Table A.8 and a format example is provided in Fig. A.9.

SINEX Troposphere Format. The SINEX troposphere format [A.36] extends the SINEX format to capture tropospheric observations and estimates. The format can contain the total and wet zenith path delays, precipitable water vapor, and gradients thereof along with the respective standard deviations. In addition, standard atmospheric measurements such as barometric pressure, temperature, and relative humidity can be provided. GNSS users can use the SINEX troposphere data to correct the range observations to improve the position estimate. Weather reporting and prediction agencies also use the observed and estimated meteorological data contained in the SINEX troposphere file for numerical modeling and climatological archives.

SINEX troposphere files follow the overall conventions of a standard SINEX file format and inherit various of its data blocks, but are identified by a `%=TRO` header line. A list of labels specific to the troposphere format is provided in Table A.9.

Bias SINEX Format. Following the example of the SINEX troposphere format, a tailored SINEX version has also been proposed for the exchange of code and phase biases of satellites and GNSS monitoring stations [A.37]. The format is specifically designed to handle differential code biases (DCB), but can also handle carrier-phase and intersystem biases. A list of associated labels for the bias-specific data blocks is provided in Table A.9. SINEX bias files are distinguished from standard SINEX files through the `%=BIA` header line.

A.2.4 IONEX Format

The IONosphere EXchange (IONEX) format [A.38] has been developed for the exchange of global ionosphere maps (GIMs) derived from the analysis of GNSS observations. Aside from space weather monitoring and ionospheric analyses, these maps enable more accurate and precise single-frequency position estimates than the Klobuchar or NeQuick models (Chap. 6) used for real-time ionospheric correction inside a GNSS receiver.

GIM products in IONEX format are generated by various IGS analysis centers [A.39] and provide estimated vertical total electron content (VTEC) values on an equidistant grid in geocentric longitude and latitude at discrete time intervals (Fig. A.10). These are used in combination with a single-layer model (Chaps. 6 and 19) to compute the ionospheric path delays in the processing of pseudorange and phase observations.


```

%=SNX2.01COD14:019:09616IGS14:013:0000014:016:00000P008481SEA
*-----
+FILE/REFERENCE
  DESCRIPTIO      CODE, Astronomical Institute, University of Bern
  OUTPU           CODE IGS 3-day solution
  INPU            CODE IGS 1-day solutions
-FILE/REFERENCE
*-----
+SITE/ID
*CODE PT  DOMES  T STATION DESCRIPTION  APPROX_LON  APPROX_LAT  APP_H
ABMF  A  97103M001 P LesAbymes, FR      298 28 20.9  16 15 44.3 -25.6
ZIMJ  A  14001M006 P Zimmerwald, CH      7 27 54.4  46 52 37.7  954.3
-SITE/ID
+SITE/RECEIVER
*SITE PT SOLN T DATA_START DATA_END DESCRIPTION S/N FIRMWARE
ABMF  A  1 P 14:013:00000 14:015:86370 TRIMBLE NETR9 -----
ZIMJ  A  1 P 14:013:00000 14:015:86370 JAVAD TRE_G3TH DELTA -----
-SITE/RECEIVER
+SITE/ANTENNA
*SITE PT SOLN T DATA_START DATA_END DESCRIPTION S/N
ABMF  A  1 P 14:013:00000 14:015:86370 TRM57971.00 NONE -----
ZIMJ  A  1 P 14:013:00000 14:015:86370 JAVRINGANT_DM NONE -----
-SITE/ANTENNA
+SITE/GPS_PHASE_CENTER
*DESCRIPTION S/N L1->ARP(M) (U,E,N) L2->ARP(M) (U,E,N)
JAVRINGANT_DM NONE ----- 0.0893 0.0011 0.0009 0.1196 0.0003 -.0001 IGS08_1771
TRM57971.00 NONE ----- 0.0668 0.0011 -.0003 0.0578 0.0001 0.0007 IGS08_1771
-SITE/GPS_PHASE_CENTER
+SITE/ECCENTRICITY
*SITE PT SOLN T DATA_START DATA_END AXE ARP->BENCHMARK(M)
ABMF  A  1 P 14:013:00000 14:015:86370 UNE 0.0000 0.0000 0.0000
ZIMJ  A  1 P 14:013:00000 14:015:86370 UNE 0.0770 0.0000 0.0000
-SITE/ECCENTRICITY
*-----
+SATELLITE/ID
*SITE PR COSPAR T DATA_START DATA_END ANTENNA
G063 01 2011-036A P 11:197:00000 00:000:00000 BLOCK IIF
R735 24 2010-007B P 10:060:00000 00:000:00000 GLONASS-M
-SATELLITE/ID
+SATELLITE/PHASE_CENTER
*SITE L SATELLITE SATELLITE_X SATELLITE_Y L SATELLITE_X SATELLITE_Y MODEL T M
G063 1 1.5613 0.3940 0.0000 2 1.5613 0.3940 0.0000 IGS08_1771 A E
R735 1 2.4830 -.5450 0.0000 2 2.4830 -.5450 0.0000 IGS08_1771 A E
-SATELLITE/PHASE_CENTER
*-----
+SOLUTION/EPOCHS
*CODE PT SOLN T DATA_START DATA_END MEAN EPOCH
ABMF  A  1 P 14:013:00000 14:015:86370 14:014:43185
ZIMJ  A  1 P 14:013:00000 14:015:86370 14:014:43185
-SOLUTION/EPOCHS
+SOLUTION/ESTIMATE
*INDEX TYPE CODE PT SOLN REF_EPOCH UNIT S ESTIMATED VALUE STD_DEV
1 STAX ABMF A 1 14:014:43200 m 2 0.291978574407741E+07 .346952E-03
2 STAY ABMF A 1 14:014:43200 m 2 -.538374500399109E+07 .547673E-03
3 STAZ ABMF A 1 14:014:43200 m 2 0.177460477599520E+07 .307686E-03
769 STAX ZIMJ A 1 14:014:43200 m 1 0.433129380015458E+07 .353701E-03
770 STAY ZIMJ A 1 14:014:43200 m 1 0.567542294104827E+06 .204759E-03
771 STAZ ZIMJ A 1 14:014:43200 m 1 0.463313582612440E+07 .365545E-03
778 XPO ---- -- 1 14:013:00000 mas 2 0.301765028159985E+02 .775857E-02
782 YPO ---- -- 1 14:013:00000 mas 2 0.329956712999544E+03 .781743E-02
786 UT ---- -- 1 14:013:00000 ms 2 -.110025006935236E+03 .145173E-03
793 SATELLITE_X G063 LC ---- 14:014:43185 m 2 0.156129999660433E+01 .145173E-05
848 SATELLITE_X R735 LC ---- 14:014:43185 m 2 0.248300000792411E+01 .145173E-05
-SOLUTION/ESTIMATE
*-----
%ENDSNX

```

Fig. A.9 SINEX format example. Labels as well as header, trailer and comment lines are highlighted in color

Table A.8 Common SINEX data blocks; (m), (r), and (o) indicate mandatory, recommended and optional blocks

Label		Contents
FILE/REFERENCE	(m)	Information on organization, software and hardware used in the file generation
FILE/COMMENT	(o)	General comments on the SINEX file
INPUT/HISTORY	(r)	Information on the agency, source data period, techniques, parameters, and content type (station, orbit, EOP, troposphere, etc.)
INPUT/FILES	(o)	Source data files
INPUT/ACKNOWLEDGEMENTS	(o)	List of contributing agencies
NUTATION/DATA	(m, VLBI)	Employed nutation model (IAU1980, IERS1996, IAU2000a/b)
PRECESSION/DATA	(m, VLBI)	Employed precession model (IAU1976, IER1996)
SOURCE/ID	(m, VLBI)	Designation of VLBI radio sources
SITE/ID	(m)	Description of key site parameters (identifiers, observation techniques, description, approximate location)
SITE/RECEIVER	(m)	Employed receivers (type, serial number, timespan) for each site
SITE/ANTENNA	(m)	Employed antennas and radomes (type, serial number, timespan) for each site
SITE/GPS_PHASE_CENTER	(m)	GPS L1/L2 phase center offsets of each antenna type
SITE/GAL_PHASE_CENTER		Galileo E1/E5a/E6/E5b/E5ab phase center offsets of each antenna type
SITE/ECCENTRICITY	(m)	Eccentricities, i. e., distances of antenna reference point from surveyed marker for each site
SATELLITE/ID	(r)	List of employed GNSS satellites with antenna type (or block), space vehicle and COSPAR number as well as associated RINEX/SP3 satellite identifier (=PRN/slot number) and period of use/assignment
SATELLITE/PHASE_CENTER	(m)	GNSS satellite antenna phase center offsets from the center-of-mass for each frequency band
SOLUTION/EPOCHS	(m)	List of observation timespan for each solution, site and point for which parameters have been estimated
BIAS/EPOCHS	(r/m)	Type (range, time, scale, etc.) and period of biases estimated in individual solutions
SOLUTION/STATISTICS	(o)	Statistical properties (no. of observations and unknowns, sampling, residuals, etc.)
SOLUTION/ESTIMATE	(m)	Estimated values and standard deviations of all solution parameters
SOLUTION/APRIORI	(r/m)	A priori values and constraints applied in the estimation
SOLUTION/MATRIX_ESTIMATE	(r/m)	Upper or lower triangle of correlation, covariance, or information matrix
SOLUTION/MATRIX_APRIORI	(m)	A priori correlation, covariance, or information matrix in triangular form
SOLUTION/NORMAL_EQUATION_VECTOR	(m)	Right-hand side of the unconstrained (reduced) normal equation
SOLUTION/NORMAL_EQUATION_MATRIX	(m)	Unconstrained normal equations matrix in triangular form

Table A.9 Data blocks specific to the SINEX Troposphere and SINEX Bias formats; (m) indicates mandatory blocks

Label		Contents
TROP/DESCRIPTION	(m)	Values of analysis parameters (sampling interval, elevation cutoff, mapping function, etc.) and list of parameters provided in the solution (estimated zenith delays, meteorological data, etc.)
TROP/STA_COORDINATES	(m)	Employed stations and their coordinates
TROP/SOLUTION	(m)	Values of estimated parameters at discrete time steps for each site
BIAS/DESCRIPTION	(m)	Specification of solution parameters (sampling interval, etc.)
BIAS/SOLUTION	(m)	Type of bias (signals, satellite, station), period of applicability, value and standard deviation

Format Description. The IONEX file format follows the RINEX2 template with 80 column records. It contains a file header followed by a TEC map for each epoch. Each TEC map is itself made up of multiple records providing TEC values for a given latitude over the specified longitude grid (Fig. A.11). Optionally,

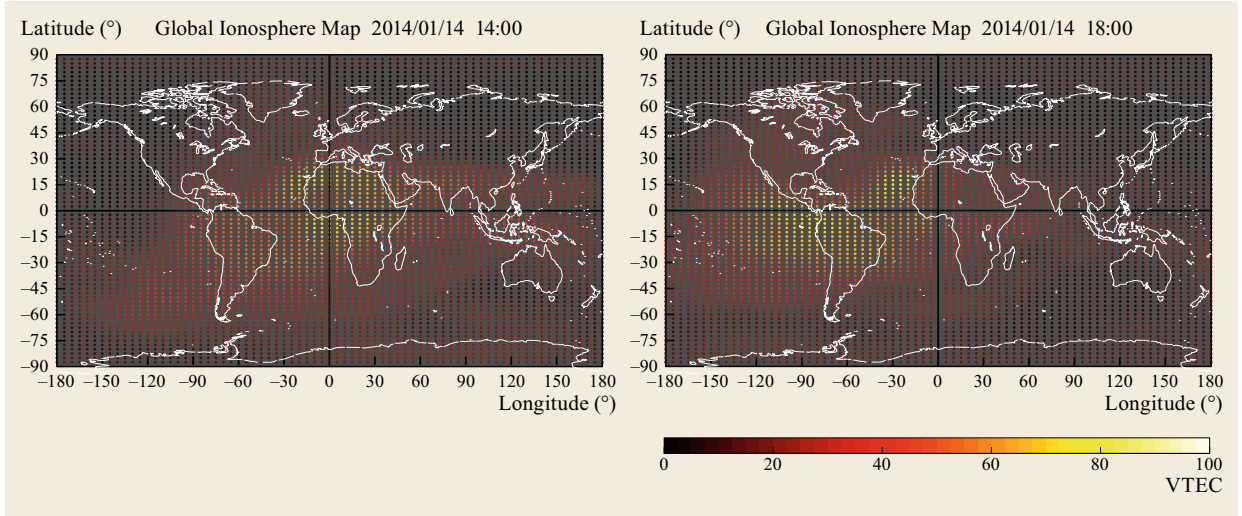


Fig. A.10 Global ionosphere maps of the IGS CODE analysis center

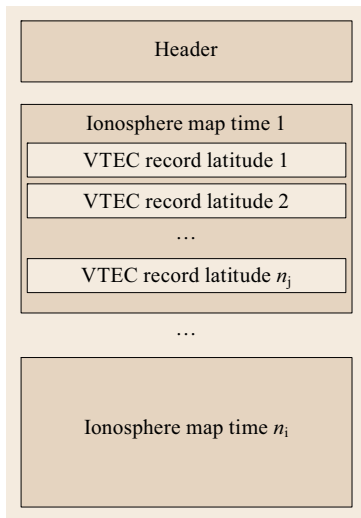


Fig. A.11 IONEX file structure

the TEC maps may be complemented by RMS TEC maps providing the expected uncertainty of the TEC values.

A truncated example illustrating the basic format of the IONEX header and data records is shown in Fig. A.12. The header primarily specifies the range in time, longitude and latitude as well as the corresponding step sizes for the TEC information given in the various TEC maps. Even though the format is designed to support three-dimensional maps, current products are confined to two-dimensional single-layer representation with a fixed height relative to a specified mean Earth radius. To facilitate interpolation, both 00:00 h and 24:00 h epochs are usually included in the daily IONEX

products. Likewise, the longitude grid for global maps includes TEC values for $\lambda = -180^\circ$ and $\lambda = +180^\circ$ despite the resulting redundancy. A second set of header lines provides the set of differential code biases (DCBs) of all satellites and stations incorporated in the generation of the TEC maps. Since ionospheric path delays are retrieved from differences of observations made at two different frequencies (Chap. 39) such biases have a direct impact on the estimated ionospheric activity. Provision of the biases in the file header ensures full transparency and consistency, even though the values are not actually required, to use TEC maps in single-frequency positioning applications.

Each TEC map is made up of a series of consecutive records providing the TEC values over the full set of longitude grid points at a specified latitude. Using a predefined scale of typically 0.1 TECU, all values can be represented through integer numbers with at most five digits. Individual TEC maps are embedded in a START OF TEC MAP and END OF TEC MAP frame and marked by their sequence number. Furthermore, the calendar date and time of the respective epoch are provided to ease their identification.

Interpolation. The global ionosphere maps provide the values $VTEC_{i,j,k}$ of the vertical total electron contents at discrete times t_i , geocentric latitudes φ'_j , and longitudes λ_k (Fig. A.10). Application of the single-layer model requires interpolation of these values for a given time t and location (λ, φ') of the ionospheric pierce point (i.e., the point at which the signal path intersects the thin shell used to represent the ionosphere; see Chap. 19, Fig. 19.1).

-----1 0-----2 0-----3 0-----4 0-----5 0-----6 0-----7 0-----8															
1.0	IONOSPHERE MAPS					GNSS		IONEX VERSION / TYPE							
ADDNEQ2 V5.3	AIUB					18-JAN-14 21:04		PGM / RUN BY / DATE							
CODE'S GLOBAL IONOSPHERE MAPS FOR DAY 014, 2014					COMMENT										
Web page: www.aiub.unibe.ch/content/ionosphere/					DESCRIPTION										
2014	1	14	0	0	0	EPOCH OF FIRST MAP									
2014	1	15	0	0	0	EPOCH OF LAST MAP									
7200						INTERVAL									
13						# OF MAPS IN FILE									
NONE						MAPPING FUNCTION									
10.0						ELEVATION CUTOFF									
One-way carrier phase leveled to code					OBSERVABLES USED										
278						# OF STATIONS									
56						# OF SATELLITES									
6371.0						BASE RADIUS									
2						MAP DIMENSION									
450.0 450.0 0.0						HGT1 / HGT2 / DHGT									
87.5 -87.5 -2.5						LAT1 / LAT2 / DLAT									
-180.0 180.0 5.0						LON1 / LON2 / DLON									
-1						EXPONENT									
TEC/RMS values in 0.1 TECU; 9999, if no value available					COMMENT										
DIFFERENTIAL CODE BIASES					START OF AUX DATA										
G01	-10.591	0.007				PRN / BIAS / RMS									

R	ZIMJ	14001M006	-13.992	0.038	STATION / BIAS / RMS										
DCB values in ns; zero-mean condition wrt satellite values					COMMENT										
DIFFERENTIAL CODE BIASES					END OF AUX DATA										
1										END OF HEADER					
2014	1	14	0	0	START OF TEC MAP										
87.5-180.0 180.0	5.0	450.0	EPOCH OF CURRENT MAP												
23	23	24	25	26	27	28	29	30	31	32	33	34	34	35	36
36	37	37	38	38	38	38	38	37	37	37	36	36	36	35	35
34	34	34	33	33	33	33	33	32	32	32	32	32	32	32	31
31	31	30	30	29	29	28	28	27	26	26	25	24	24	23	23
22	22	22	22	22	22	22	22	22	23						
-----										Omitted lines					
-87.5-180.0 180.0	5.0	450.0	LAT/LON1/LON2/DLON/H												
190	190	191	191	191	191	190	190	189	188	188	187	186	185	184	183
182	181	179	178	177	176	175	174	173	171	170	169	168	167	165	164
163	162	161	160	158	157	157	156	155	154	154	154	153	153	154	154
154	155	156	157	158	159	161	163	164	166	168	170	172	174	176	178
180	182	183	185	186	187	188	189	190							
1										END OF TEC MAP					
-----										Omitted lines					
										END OF FILE ...					
-----1 0-----2 0-----3 0-----4 0-----5 0-----6 0-----7 0-----8										}					

Fig. A.12 IONEX file format example. Rulers at the start and end have been added for information only and are not part of the actual format

For interpolation to the given location a bilinear interpolation

$$\begin{aligned} \text{VTEC}_i(\lambda, \varphi') &= (1-p)(1-q)\text{VTEC}_{i,j,k} \\ &+ (p)(1-q)\text{VTEC}_{i,j+1,k} \\ &+ (1-p)(q)\text{VTEC}_{i,j,k+1} \\ &+ (p)(q)\text{VTEC}_{i,j+1,k+1} \end{aligned} \quad (\text{A.7})$$

with coefficients

$$\begin{aligned} p &= (\varphi' - \varphi_j') / (\varphi_{j+1}' - \varphi_j') \\ q &= (\lambda - \lambda_k) / (\lambda_{k+1} - \lambda_k) \end{aligned} \quad (\text{A.8})$$

is applied across the intervals $\varphi_j' \leq \varphi' < \varphi_{j+1}'$ and $\lambda_k \leq \lambda < \lambda_{k+1}$, limited by the surrounding grid points.

For interpolation in time, a linear interpolation

$$\begin{aligned} \text{VTEC}(t, \lambda, \varphi') &= (1-\tau)\text{VTEC}_i(\lambda, \varphi') \\ &+ \tau\text{VTEC}_{i+1}(\lambda, \varphi') \end{aligned} \quad (\text{A.9})$$

with $\tau = (t - t_i) / (t_{i+1} - t_i)$ can subsequently be employed across the time interval $t_i \leq t < t_{i+1}$. Improved results may, however be obtained, by taking into account that ionospheric activity varies mostly with local time rather than UTC. This results in an apparent westwards shift of the average electron density distribution at a rate of $\omega = 15^\circ/\text{h}$, which is illustrated in Fig. A.10. A modified relation

$$\begin{aligned} \text{VTEC}(t, \lambda, \varphi') &= (1-\tau)\text{VTEC}_i(\lambda, \varphi') \\ &+ \tau\text{VTEC}_{i+1}(\lambda - \omega(t_{i+1} - t_i), \varphi') \end{aligned} \quad (\text{A.10})$$

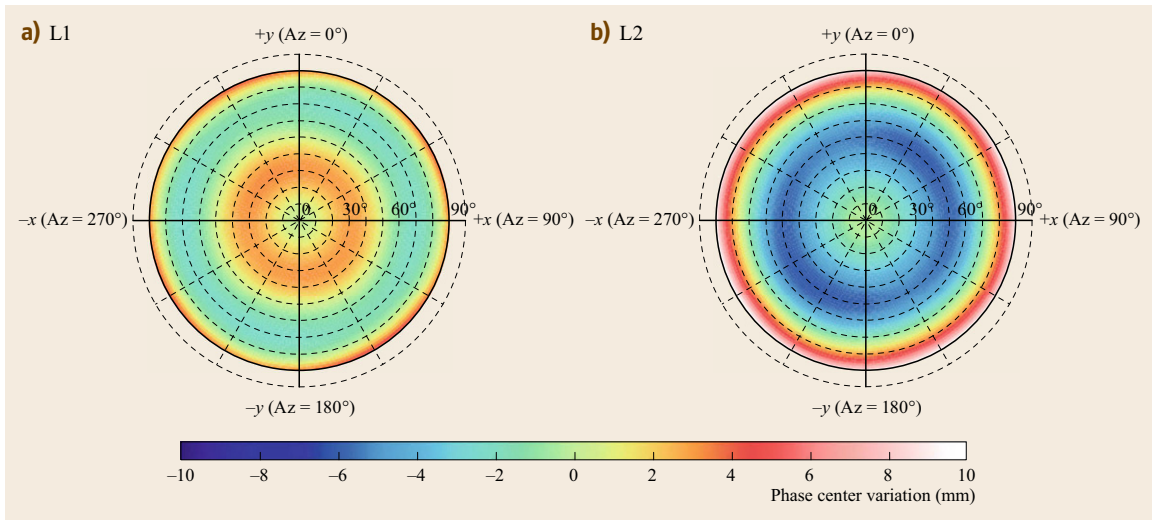


Fig. A.13 Azimuth and boresight angle-dependent phase center variation of Leica AR25.R4 antenna as provided in the `igs08.atx` antenna calibration file for GPS L1 (a) and L2 (b) signals. Image courtesy of A. Hauschild

is therefore recommended by [A.38] for interpolation of IONEX TEC maps in time. Here, the second map is shifted in longitude with respect to the first one to compensate for Earth rotation between the respective epochs.

A.2.5 ANTEX

The ANTenna EXchange (ANTEX) format [A.40] was developed to facilitate the documentation and distribution of phase center offsets (PCOs) and phase center variations (PCVs) for GNSS receiver and satellite antennas (Fig. A.13). These corrections are used in GNSS precise point positioning applications as well as GNSS satellite orbit and clock determination for high-accuracy modeling of carrier-phase observations (Chap. 19).

Except for PCV data that require an extended line width, ANTEX files employ a line format with parameters in columns 1–60 and descriptive labels in columns 61–80. Following a brief header, a series of data records with information for individual antennas are provided (Fig. A.14). Each record is itself made up of an antenna-specific header and several sets of PCO/PCV data for distinct constellations and frequency bands.

A truncated format example illustrating the basic structure of the ANTEX file and the data blocks is provided in Fig. A.15. Different colors are used to highlight the header of individual antenna records (blue) as well as PCO/PCV data for the GPS L1 (red) and L2 (green) frequency. For GNSS satellite antennas, the antenna type (associated with the block type of a GPS satellite), the three-character RINEX satellite identifier (constellation letter and PRN or slot number), the space

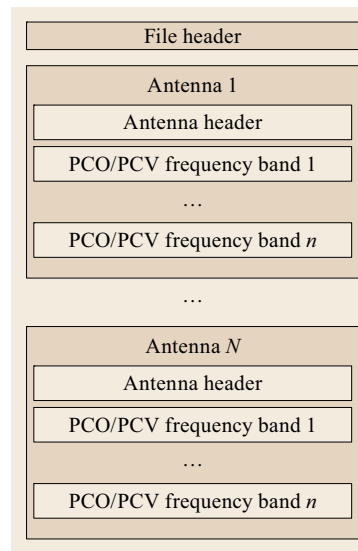


Fig. A.14 Basic structure of ANTEX antenna data files

vehicle number and the international (COSPAR) satellite number are specified in the first record header line. Next, the grid of azimuth and boresight angle values for the provision of phase center variations is specified. In addition, a validity period is specified that reflects the timespan during which the space vehicle was assigned the given satellite identifier. For GNSS receiver antennas, the antenna and radome name are identified along with the method used for calibration of the phase pattern. Since 2005, absolute antenna phase patterns are exclusively used within the IGS [A.41]. These may be based on either robot calibrations or anechoic chamber measurements.

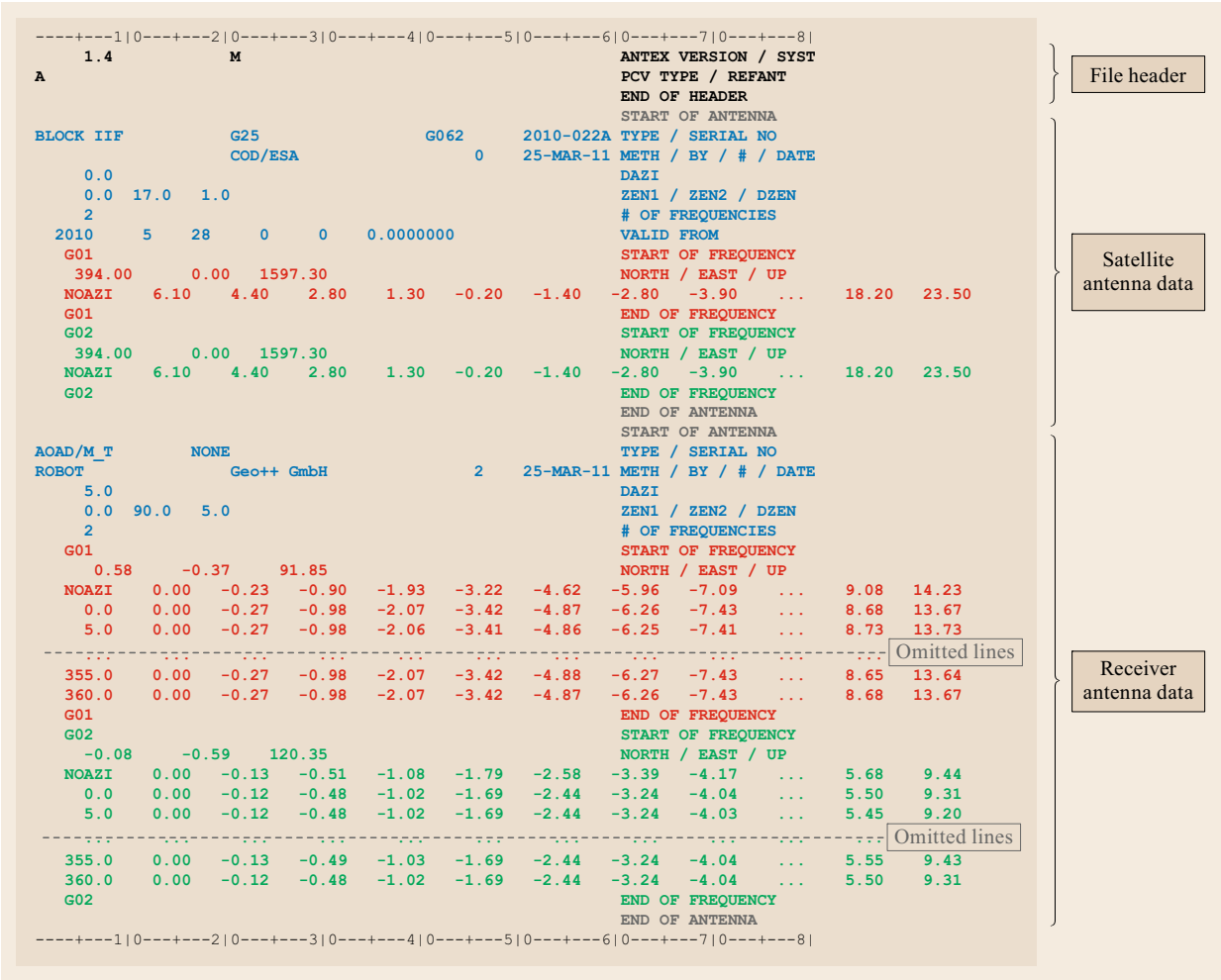


Fig. A.15 ANTEX file format example. Rulers at the start and end have been added for information only and are not part of the actual format. Ellipses denote characters or lines that have been deleted to fit the available print space

Antenna information for a specific frequency band is embedded in a START OF FREQUENCY ...END OF FREQUENCY block identified by the RINEX constellation letter and frequency band number. The offset of the antenna phase center is first provided, thereafter the phase center variations are given. Receiver antenna phase center offsets are defined with respect to the antenna reference point (ARP) and a coordinate system aligned with the nominal North, East, and Up directions. GNSS satellite phase center offsets are defined with respect to the satellites center-of-mass and body-fixed x-, y- and z-coordinates. Phase center variations referring to the same reference system are provided in the form of an averaged boresight angle-dependent pattern (marked by the NOAZI keyword) as well as an optional azimuth and boresight angle-dependent pattern. For GNSS satellites only the first

form is presently provided in the standard IGS ANTEX product, although various efforts have already been made to derive two-dimensional PCV maps for all GPS and GLONASS satellites. For a given boresight direction, the PCV maps can be interpolated using either linear (for one-dimensional PCVs) or bilinear interpolation (for two-dimensional PCVs).

A.2.6 Site Log Format

Building and maintaining a GNSS station or network of stations requires a standardized station information archive that describes the station including location, equipment, responsible agency and contact information. To meet these needs, the IGS has established a dedicated site log format. Site logs are stored as text files with an 80-character width and made up of 14 numbered sections described in Table A.10. Indi-

```

KZN2 Site Information Form
International GNSS Service

0. Form
  Prepared by (full name) : Renat Zagretdinov
  Date Prepared          : 2012-02-24
  Report Type           : NEW

1. Site Identification of the GNSS Monument
  Site Name              : KAZAN
  Four Character ID      : KZN2
  Monument Inscription   : KFU GNSS STATION
  IERS DOMES Number      : 12374M001
  CDP Number             : NONE
  Monument Description   : PILLAR
    Height of the Monument : 13(m)
    Monument Foundation    : CONCRETE BLOCK
    Foundation Depth       : 2(m)
  Marker Description     : (CHISELLED CROSS/DIVOT/BRASS NAIL/etc)
  Date Installed         : 2010-10-01
  Geologic Characteristic : CLAY and SAND
  ...

2. Site Location Information
  City or Town           : KAZAN
  State or Province      : TATARSTAN
  Country                : Russian Federation
  Tectonic Plate         : Eurasian
  Approximate Position (ITRF)
    X coordinate (m)     : 2352345.7
    Y coordinate (m)     : 2717466.1
    Z coordinate (m)     : 5251458.5
    Latitude (N is +)    : +554726.82
    Longitude (E is +)   : +0490709.28
    Elevation (m,ellips.) : 94.6
  Additional Information :

3. GNSS Receiver Information
3.1 Receiver Type       : TRIMBLE NETR9
  Satellite System      : GPS+GLO+GAL+CMP
  Serial Number         : 5049K72275
  Firmware Version      : 4.43
  Elevation Cutoff Setting : 5
  Date Installed        : 2012-02-24T00:00Z
  Date Removed          : 2012-08-14T13:00Z
  Temperature Stabiliz. : 20-30
  ...

4. GNSS Antenna Information
4.1 Antenna Type        : TRM59800.00
  Serial Number          : 5106354023
  Antenna Reference Point : BPA
  Marker->ARP Up Ecc. (m) : 0.0750
  Marker->ARP North Ecc (m) :
  Marker->ARP East Ecc (m) :
  Alignment from True N   : 0
  Antenna Radome Type     : SCIS
  Radome Serial Number    : 0702
  Antenna Cable Type      : LMR400, Times Microwave Systems
  Antenna Cable Length    : 30(m)
  Date Installed          : 2012-02-24T00:00Z
  Date Removed            : (CCYY-MM-DDThh:mmZ)
  ...

```

Fig. A.16 Site log format example (truncated sections are indicated by ellipses)

Table A.10 Sitelog file contents

No.	Section	Contents
0	Form	Lists the author, preparation date, type (new or update), link to previous site log and list of changes
1	Site identification	Site name, four-character ID, monument inscription, IERS domes number, CDP number; monument description, height (m) and foundation (type and depth); marker description and date installed, geological characteristics of foundation (soil, rock)
2	Site location	City or town, state, country, tectonic plate, approximate Cartesian and geographic coordinates
3	GNSS receiver	Receiver type, GNSS systems supported, serial number, firmware version, elevation cutoff angle, date (installed and removed). Repeated for each change
4	GNSS antenna	Antenna type, serial number, antenna reference point (ARP), marker-to-antenna ARP offset, antenna alignment w.r.t true north, radome type and serial number, antenna cable type and length, date (installed and removed). Repeated for each change
5	Surveyed local ties	Tied marker: name, usage (SLR, VLBI, control), CDP number, domes number, differential components (ITRS) (m) : dx, dy and dz, accuracy, survey method, date and additional information as required. Repeated for each additional tie and campaign
6	Frequency standard	Type (internal, external and type), input frequency, effective dates beginning and end. Repeated for each change
7	Collocation information	List of instrumentation present an the site: type (DORIS, SLR, VLBI, etc.), status, effective dates (start and end) and notes
8	Meteorological instrumentation	Humidity sensor: manufacturer, serial number, sampling interval, accuracy, aspiration, height difference to antenna, calibration date and start and end date. Corresponding information for pressure sensor, temperature sensor, water vapor radiometer and other meteorological sensors
9	Local conditions	Radio interference, multipath sources, signal obstruction
10	Local episodic effects	Date, event (tree clearing, construction, etc.)
11	On site contact	Agency, abbreviation, address, primary contact: name, telephone, fax and email, repeated for secondary contact and additional information
12	Responsible agency	Same content as in site contact section
13	More information	Primary and secondary data center, URL for station information, availability of site map, diagram, horizon mask, detailed monument description and pictures, antenna graphics and additional information

vidual parameters within each section are identified by predefined labels with colons in column 31 separating the labels and parameters. When equipment or site conditions change the relevant block is updated.

In this way, site logs provide users with a complete record of all the changes that have taken place. A truncated format example of a site log file is shown in Fig. A.16.

References

A.1

J. Januszewski: Satellite navigation systems, data messages, data transfer and formats, 11th Int. Conf. Transp. Syst. Telemat., TST 2011, Katowice-Ustrón, 2011, Communications in Computer and Information Science, Vol. 239, ed. by J. Mikulski (Springer, Berlin, Heidelberg 2011) pp. 338–345

A.2

National Marine Electronics Association: <http://www.nmea.org>

A.3

NMEA 0183 Interface Standard, v4.10 (National Marine Electronics Association, Severna Park 2011)

A.4

W. Gurtner, G. Mader, D. MacArthur: A common exchange format for GPS data, CIGNET Bull. 2(3), 1–11 (1989)

A.5

W. Gurtner: RINEX: The receiver-independent exchange format, GPS World 5(7), 48–52 (1994)

A.6

W. Gurtner, L. Estey: RINEX: The Receiver Independent Exchange Format – Version 2.11, 26 Jun. 2012 (IGS/RTCM RINEX Working Group, 2012)

A.7

RINEX – The Receiver Independent Exchange Format – Version 3.03, 14 July 2015 (IGS RINEX WG, RTCM-SC104, Pasadena 2015)

A.8

International GNSS Service: IGS Formats <http://kb.igs.org/hc/en-us/articles/201096516-IGS-Formats>

A.9

Y. Hatanaka: A compression format and tools for GNSS observation data, Bull. Geogr. Surv. Inst. 55, 21–29 (2008)

A.10

Navstar GPS Space Segment/Navigation User Interfaces, Interface Specification, IS-GPS-200H, 24 Sep. 2013 (Global Positioning Systems Directorate, Los Angeles 2013)

- A.11 Quasi-Zenith Satellite System Interface Specification – Satellite Positioning, Navigation and Timing Service, IS-QZSS-PNT-001, Draft 12 July 2016 (Cabinet Office, Tokyo 2016)
- A.12 BeiDou Navigation Satellite System Signal In Space Interface Control Document – Open Service Signal, v2.1, Nov. 2016 (China Satellite Navigation Office, Beijing 2013)
- A.13 European GNSS Service Centre: European GNSS (Galileo) Open Service Signal In Space Interface Control Document, OS SIS ICD, Iss. 1.3, Dec. 2016 (EU 2016)
- A.14 Indian Regional Navigation Satellite System – Signal In Space ICD for Standard Positioning Service, version 1.0, June 2014 (Indian Space Research Organization, Bangalore 2014)
- A.15 Global Navigation Satellite System GLONASS – Interface Control Document, v5.1, (Russian Institute of Space Device Engineering, Moscow 2008)
- A.16 G. Weber, D. Dettmering, H. Gebhard, R. Kalafus: Networked transport of RTCM via internet protocol (Ntrip) – IP-streaming for real-time GNSS applications, ION GPS 2005, Long Beach, 2005 (ION, 2005) pp. 2243–2247
- A.17 RTCM 10403.3, Differential GNSS Services, Version 3, 7 Oct. 2016 (RTCM, Arlington)
- A.18 Radio Technical Commission for Maritime Services: <http://www.rtcn.org>
- A.19 A. Boriskin, D. Kozlov, G. Zyryanov: The RTCM multiple signal messages: A new step in GNSS data standardization, ION GNSS 2012, Nashville, 2012 (ION, 2012) pp. 2947–2955
- A.20 M. Schmitz: RTCM state space representation messages, status and plans, PPP-RTK & Open Stand. Symp., Frankfurt, 2012 (BKG, Frankfurt a.M. 2012) pp. 1–31
- A.21 L. Estey, C. Meertens, D. Mencin: Application of BINEX and TEQC for real-time data management, IGS Netw. Workshop, Oslo, 2000 (IGS, Pasadena 2000)
- A.22 L. Estey, D. Mencin: BINEX as a format for near-real time GNSS and other data streams, G43A-0663, AGU Fall Meet. 2008, San Francisco, 2008 (AGU, Washington D.C. 2008)
- A.23 UNAVCO: BINEX: Binary Exchange Format, <http://binex.unavco.org/binex.html>
- A.24 B. Remondi: Extending the National Geodetic Survey Standard for GPS Orbit Formats, NOAA Technical Report NOS 133 NGS 46 (National Geodetic Information Branch, NOAA, Rockville, MD, Nov. 1989)
- A.25 S. Hilla: Extending the standard product 3 (SP3) orbit format, International GPS Serv. Netw. Data Anal. Cent. Workshop, Ottawa, 2002 (IGS, Pasadena 2002)
- A.26 M. Schenewerk: A brief review of basic GPS orbit interpolation strategies, GPS Solut. **6**(4), 265–267 (2003)
- A.27 H. Yousif, A. El-Rabbany: Assessment of several interpolation methods for precise GPS orbit, J. Navig. **60**(3), 443–455 (2007)
- A.28 M. Horemuz, J.V. Andersson: Polynomial interpolation of GPS satellite coordinates, GPS Solut. **10**(1), 67–72 (2006)
- A.29 W.H. Press, S.A. Teukolsky, W.T. Vetterling: *Numerical Recipes: The Art of Scientific Computing* (Cambridge Univ. Press, Cambridge 2007)
- A.30 B. Hofmann-Wellenhof, H. Lichtenegger, E. Wasle: *GNSS: Global Navigation Satellite Systems: GPS, GLONASS, Galileo, and More* (Springer, Berlin, Heidelberg 2008)
- A.31 G. Xu: *Orbits* (Springer, Berlin, Heidelberg 2008)
- A.32 J.F. Zumberge, G. Gendt: The demise of selective availability and implications for the international GPS service, Phys. Chem. Earth (A) **26**(6–8), 637–644 (2001)
- A.33 J. Ray, W. Gurtner: RINEX Extensions to Handle Clock Information – Version 3.02, 2 Sep. 2010
- A.34 IGS: SINEX – Solution (Software/technique) INdependent EXchange Format, Version 2.02, 1 Dec. 2006, <https://www.iers.org/IGS/EN/Organization/AnalysisCoordinator/SinexFormat/sinex.html>
- A.35 M. Rothacher, H. Drewes, A. Nothnagel, B. Richter: Integration of space geodetic techniques as the basis for a global geodetic-geophysical observing system (GGOS-D): An overview. In: *System Earth via Geodetic-Geophysical Space Techniques* (Springer, Berlin 2010) pp. 529–537
- A.36 IGS: SINEX_TRO – Solution (Software/technique) INdependent EXchange Format for combination of TROpospheric estimates, Version 0.01, 01 Mar. 1997, ftp://igs.org/pub/data/format/sinex_tropo.txt
- A.37 SINEX_BIAS – Solution (Software/technique) INdependent EXchange Format for GNSS Biases, Version 1.00, 7 Dec 2016, <http://www.biasws2015.unibe.ch/documents.html>
- A.38 S. Schaer, W. Gurtner, J. Feltens: IONEX: The IONosphere map EXchange format Version 1, IGS AC Workshop, Darmstadt, 1998 (IGS, Pasadena 1998)
- A.39 M. Hernández-Pajares, J.M. Juan, J. Sanz, R. Orus, A. García-Rigo, J. Feltens, A. Komjathy, S.C. Schaer, A. Krankowski: The IGS VTEC maps: a reliable source of ionospheric information since 1998, J. Geodesy **83**(3–4), 263–275 (2009)
- A.40 M. Rothacher, R. Schmid: ANTEX: The Antenna Exchange Format, Version 1.4, 15 Sep. 2010, <ftp://igs.org/pub/station/general/antex14.txt>
- A.41 R. Schmid, M. Rothacher, D. Thaller, P. Steigenberger: Absolute phase center corrections of satellite and receiver antennas, GPS Solut. **9**(4), 283–293 (2005)

Annex B: GNSS Parameters

Oliver Montenbruck, Michael Meurer, Peter Steigenberger

This chapter summarizes important GNSS-related parameters. Next to general physical constants, key parameters of the GNSS constellations and the various GNSS signals are provided.

B.1 Physical Constants

Physical constants and parameters of relevance for the generation and processing of GNSS observations are provided in Table B.1.

B.2 Orbital Parameters

Table B.2 summarizes the orbital parameters of current global and regional navigation satellite systems.

B.3 Signals

The spectra of current and planned GNSS navigation signals are illustrated in Fig. B.1. Key parameters of the individual signals are compiled in Table B.3.

Signal bandwidths (BW) given in the table refer to the location of the first minimum outside the main lobe(s). The actual spectral usage depends on the filtering and may cover a larger frequency range, particularly

when sharing a frequency band with other wideband signals. GNSS signals are commonly described as

$$S = s_I \cos(2\pi ft) \pm s_Q \sin(2\pi ft) , \quad (\text{B.1})$$

where the *cos*-component of the signal is designated as the *in-phase* component (I), while the *sin*-signal is termed the *quadrature* component (Q) [B.11]. Depending on the choice of sign in the above equation, the instantaneous phase of the Q-channel is either behind (plus-sign) or ahead (minus-sign) of the I-channel. To distinguish both options, we use the notations

$$S = s_{I^+} \cos(2\pi ft) + s_{Q^+} \sin(2\pi ft) \quad (\text{B.2})$$

and

$$S = s_{I^-} \cos(2\pi ft) - s_{Q^-} \sin(2\pi ft) , \quad (\text{B.3})$$

for the specification of the channel (Ch), where the superscript of the I- and Q-symbols indicates the employed signal description. In the absence of a superscript, the association of signals to the I- and Q-channel and/or the concise phase relation is not traceable from public information. However, signals designated as *I* and *Q*, respectively, are known to be modulated in phase-quadrature with respect to each other.

Table B.1 Physical constants and parameters

Quantity	Description	Value	Unit	References and remarks
Time				
TT–TAI	Time offset TT and TAI	32.184	s	IAU 1991 [B.1]
GPST–TAI	Time offset GPS time and TAI	≈ -19	s	[B.2]
BDT–TAI	Time offset BDS time and TAI	≈ -33	s	[B.3]
Universal				
c	Speed of light in vacuum	$2.99792458 \cdot 10^8$	m s^{-1}	Defining constant [B.4]
G	Constant of gravitation	$6.67408 \cdot 10^{-11}$	$\text{m}^3 \text{kg}^{-1} \text{s}^{-2}$	[B.4]
μ_0	Permeability of vacuum	$12.566370614 \dots \cdot 10^{-7}$	N A^{-2}	$4\pi \cdot 10^{-7}$ [B.4]
ϵ_0	Permittivity of vacuum	$8.854187817 \dots \cdot 10^{-12}$	F m^{-1}	$1/(\mu_0 c^2)$ [B.4]
e	Elementary charge	$1.6021766208 \cdot 10^{-19}$	C	[B.4]
m_e	Electron mass	$9.10938356 \cdot 10^{-31}$	kg	[B.4]
Earth				
GM_\oplus	Geocentric grav. constant	$3.986004415 \cdot 10^{14}$	$\text{m}^3 \text{s}^{-2}$	EGM2008, TT-compatible [B.5]
J_2	Dynamic form factor	$1.08263 \cdot 10^{-3}$		GRS80 [B.6]
R_\oplus	Equatorial radius	$6.378137 \cdot 10^6$	m	GRS80 [B.6]
$1/f$	Flattening factor	298.257222101		GRS80 [B.6]
ω_\oplus	Nominal mean angular velocity	$7.292115 \cdot 10^{-5}$	rad s^{-1}	GRS80 [B.6]
Sun				
GM_\odot	Heliocentric grav. constant	$1.32712440040944 \cdot 10^{20}$	$\text{m}^3 \text{s}^{-2}$	DE421 [B.7]
AU	Astronomical unit	$1.49597870700 \cdot 10^{11}$	m	[B.8]
R_\odot	Mean solar radius	$6.96 \cdot 10^8$	m	[B.9]
TSI	Total solar irradiance	1360.8	W m^{-2}	[B.10]
P_\odot	Radiation pressure at 1 AU	$4.5391 \cdot 10^{-6}$	N m^{-2}	TSI/ c
Moon				
$GM_\text{☾}$	Selenocentric grav. constant	$4.902800076 \cdot 10^{12}$	$\text{m}^3 \text{s}^{-2}$	DE421 [B.7]
$R_\text{☾}$	Mean lunar radius	$1.7374 \cdot 10^6$	m	[B.9]

Table B.2 Representative orbital parameters (period, semi-major axis a , height above the Earth h , eccentricity e , and inclination i) of global and regional navigation satellite system satellites. A period of n/m revolutions (rev) per sidereal day (d_{sid}) results in a ground-track repeat track after m inertial rotations of the Earth (approximately $m \times 23^{\text{h}}56^{\text{m}}$)

System	Period (rev/ d_{sid})	Period (h)	a (km)	h (km)	e	i (°)
GLONASS	17/8	$11^{\text{h}}16^{\text{m}}$	25 510	19 130	0.0	64.8
GPS	2/1	$11^{\text{h}}58^{\text{m}}$	26 560	20 180	0.0	55
BeiDou (MEO)	13/7	$12^{\text{h}}53^{\text{m}}$	27 910	21 530	0.0	55
Galileo	17/10	$14^{\text{h}}05^{\text{m}}$	29 600	23 220	0.0	56
QZSS (IGSO)	1/1	$23^{\text{h}}56^{\text{m}}$	42 160	35 790	0.1	43
BeiDou (IGSO)	1/1	$23^{\text{h}}56^{\text{m}}$	42 160	35 790	0.0	55
IRNSS (IGSO)	1/1	$23^{\text{h}}56^{\text{m}}$	42 160	35 790	0.0	29
BeiDou/IRNSS/QZSS (GEO), SBAS	1/1	$23^{\text{h}}56^{\text{m}}$	42 160	35 790	0.0	≤ 2

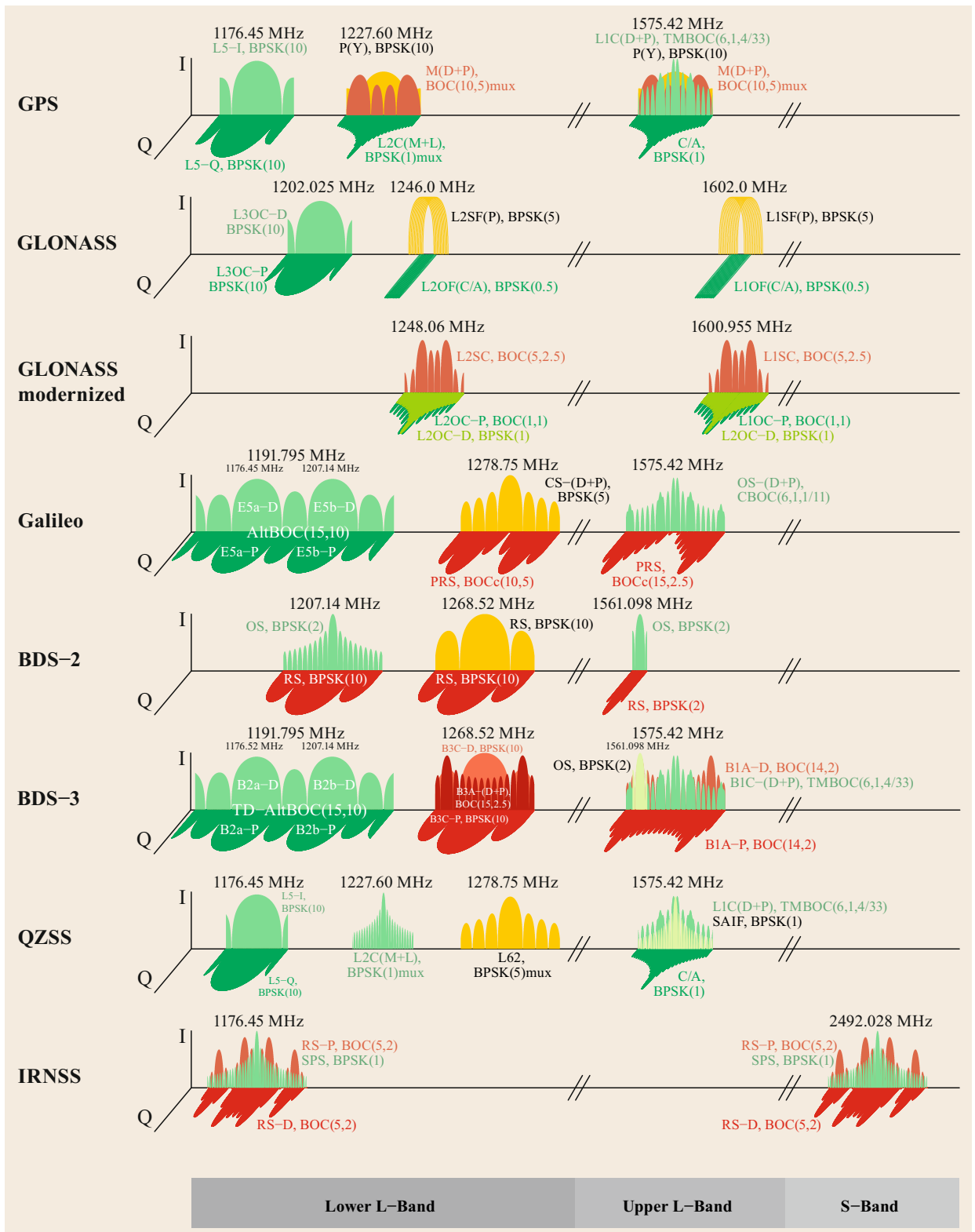


Fig. B.1 GNSS signals overview. Colors indicate open signals (shades of green), authorized signals (shades of red), and signals that can be tracked with restrictions (yellow)

Table B.3 GNSS signals overview. The specified bandwidth (BW) refers to the location of the first minimum outside the main lobe(s) of the signal and is generally smaller than the actual transmission bandwidth and onboard bandpass filtering. Question marks indicate missing or undefined information

Sys	Band	Signal	Frequency (MHz)	BW (MHz)	Ch	Modulation	Rate (MHz)	Code prim./second. (chips)	Type	Data (bps/sps)	Power (min. rcvd.) (dBW)	References
GPS	L1	P(Y)	1575.42	±10	I ⁺	BPSK(10)	10.23	$6.9 \cdot 10^{12}$	M-seq.	50/50	-161.5	[B.2, 12]
		C/A		±1	Q ⁺	BPSK(1)	1.023	1023	Gold	50/50	-158.5	[B.2, 12]
		L1C-D		±2	I ⁺	TMBOC(6,1.4/33)	1.023	10 230	Weil	50/100	-163.0	[B.13]
		L1C-P		±2	I ⁺	TMBOC(6,1.4/33)	1.023	10 230/1800	Weil	–	-158.25	[B.13]
		M-D		±15	I ⁺	BOC(10,5) mux	5.115	n/a	n/a	≤100/200	-158.0	[B.11, 14, 15]
		M-P		±15	I ⁺	BOC(10,5) mux	5.115	n/a	n/a	n/a	-158.0	[B.11, 14, 15]
		L2		±10	I ⁺	BPSK(10)	10.23	$6.9 \cdot 10^{12}$	M-seq.	50/50	-164.5 ^a , -161.5 ^{b,c,d}	[B.2, 12]
		P(Y)	1227.60		I ⁺	BPSK(10)	0.5115	10 230	M-seq.	25/50	-163.0 ^{b,c} , -161.5 ^d	[B.2, 12]
		L2 CM		±1	Q ^{+e}	BPSK(1) mux	0.5115	767 250	M-seq.	–	-163.0 ^{b,c} , -161.5 ^d	[B.2, 12]
		L2 CL		±1	Q ^{+e}	BPSK(1) mux	0.5115	n/a	n/a	≤100/200	-164.0	[B.11, 14, 15]
		M-D		±15	I ⁺	BOC(10,5) mux	5.115	n/a	n/a	n/a	-164.0	[B.11, 14, 15]
		M-P		±15	I ⁺	BOC(10,5) mux	5.115	n/a	n/a	n/a	-164.0	[B.11, 14, 15]
	L5	L5I	1176.45	±10	I ⁺	BPSK(10)	10.23	10 230/10	M-seq.	50/100	-157.9 ^e , -157.0 ^d	[B.16]
		L5Q		±10	Q ⁺	BPSK(10)	10.23	10 230/20	M-seq.	–	-157.9 ^e , -157.0 ^d	[B.16]
GLO	L1	L1SF (P)	$1602.0 + k \cdot 0.5625^f$	±5	I	BPSK(5)	5.11	5 110 000	M-seq.	50	n/a	[B.17, 18]
		L1OF (C/A)	$1602.0 + k \cdot 0.5625^f$	±0.5	Q	BPSK(0.5)	0.511	511	M-seq.	50	-161.0	[B.19]
	L2	L2SF (P)	$1246.0 + k \cdot 0.4375^f$	±5	I	BPSK(5)	5.11	5 110 000	M-seq.	50	n/a	[B.17, 18]
		L2OF (C/A)	$1246.0 + k \cdot 0.4375^f$	±0.5	Q	BPSK(0.5)	0.511	511	M-seq.	50	-161.0	[B.19]
	L3	L3OC-D	1202.025	±10	I ⁻	BPSK(10)	10.23	10 230	Kasami	100/200	?	[B.20, 21]
		L3OC-P	1202.025	±10	Q ⁻	BPSK(10)	10.23	10 230	Kasami	–	?	[B.20, 21]
	L1	L1SC ^g	1600.955	±5	I ⁻	BOC(5,2.5)	5.115	?	?	?	?	[B.21, 22]
		L1OC-D ^g	1600.955	±1	Q ⁻	BPSK(1) mux	0.5115	1023/2	Gold	125/250	?	[B.21, 22]
		L1OC-P ^g	1600.955	±2	Q ⁻	BOC(1,1) mux	0.5115	1023	Gold	–	?	[B.21, 22]
		L1OCM ^h	1575.42	?	?	?	?	?	?	?	?	[B.24]
	L2	L2SC ^g	1248.06	±7	I ⁻	BOC(5,2.5)	5.115	?	?	?	?	[B.21, 23]
		L2OC-D ^g	1248.06	±1	Q ⁻	BPSK(1) mux	?	?	?	?	?	[B.21, 23]
		L2OC-P ^g	1248.06	±2	Q ⁻	BOC(1,1) mux	0.5115	10 230/50	Kasami	–	?	[B.21, 23]
	L5	L5OCM ^h	1176.45	?	?	?	?	?	?	?	?	[B.24]
				?	?	?	?	?	?	?	?	

Abbreviations: Sys = System; BW = bandwidth; Ch = channel; mux = multiplexed; n/a = nonavailability of public information for regulated/military services
Notes: ^a Block IIA/IIR; ^b Block IIR-M; ^c Block IIF; ^d Block III; ^e Nominal phase relationship, see bit 273 of CNAV msg 10; ^f frequency channel number $k = -7 \dots +6$; ^g planned; ^h study

Table B.3 (continued)

Sys	Band	Signal	Frequency (MHz)	BW (MHz)	Ch	Modulation	Rate (MHz)	Code prim./second. (chips)	Type	Data (bps/sps)	Power (min. rcvd.) (dBW)	References
GAL	E1	OS-D(B)	1575.42	±2	I ⁻	CBOC(6,1,1/11)	1.023	4092	rand.	125/250	-160.0	[B.25,26]
		OS-P(C)		±2	I ⁻	CBOC(6,1,1/11)	1.023	4092/25	rand.	–	-160.0	[B.25,26]
		PRS(A)		±17	Q ⁻	BOC _{cos} (15,2.5)	2.5575	n/a	n/a	n/a	n/a	[B.26]
	E6	CS-D(B)	1278.75	±5	I ⁻	BPSK(5)	5.115	5115	rand.	500/1000	-158.0	[B.25]
		CS-P(C)		±5	I ⁻	BPSK(5)	5.115	5115/100	rand.	–	-158.0	[B.25]
	E5ab	PRS(A)		±15	Q ⁻	BOC _{cos} (10,5)	5.115	n/a	n/a	n/a	n/a	[B.26]
			1191.795	±25		AltBOC(15,10) ^j						[B.25]
		E5b-D	1207.14	±10	I ⁻		10.23	10230/20	M-seq.	25/50	-158.0	[B.25]
	E5b	E5b-P		±10	Q ⁻		10.23	10230/100	M-seq.	25/50	-158.0	[B.25]
		E5a-D	1176.45	±10	I ⁻		10.23	10230/20	M-seq.	25/50	-158.0	[B.25]
BDS-2	B1-2	E5a-P		±10	Q ⁻		10.23	10230/100	M-seq.	25/50	-158.0	[B.25]
		OS	1561.098	±2	I ⁺	BPSK(2)	2.046	2046 ^j , 2046/20 ^k	LFSR	250/500 ^j 25/50 ^k	-163.0	[B.3]
		RS		±2	Q ⁺	BPSK(2)	2.046	n/a	n/a	n/a	n/a	
	B3	RS	1268.52	±10	I ⁺	BPSK(10)	10.23	10230/20	LFSR	n/a	n/a	[B.27,28]
		RS		±10	Q ⁺	BPSK(10)	10.23	n/a	n/a	n/a	n/a	
BDS-3	B1-2	OS	1561.098	±2	I	BPSK(2)	2.046	2046/20 ^k	LFSR	25/50 ^k	-163.0	[B.3]
		B1C-D ^l	1575.42	±2	?	TMBOC(6,1,4/33)	1.023	10230	?	50/100	?	[B.29,30]
		B1C-P ^l		±2	?	TMBOC(6,1,4/33)	1.023	10230/20	?	–	?	[B.29,30]
	B3	B1A-D ^m		±16	?	BOC(14,2)	n/a	n/a	n/a	50/100	n/a	[B.29,30]
		B1A-P ^m		±16	?	BOC(14,2)	n/a	n/a	n/a	n/a	n/a	[B.29,30]
	B3	B3C-D ^m	1268.52	±10	?	BPSK(10)	10.23	n/a	n/a	500/500	n/a	[B.29,30]
		B3C-P ^m		±10	?	BPSK(10)	10.23	n/a	n/a	500/500	n/a	[B.29,30]
		B3A-D ^m		±17	?	BOC(15,2.5)	2.5575	n/a	n/a	n/a	n/a	[B.29,30]
	B2	B3A-P ^m		±17	?	BOC(15,2.5)	2.5575	n/a	n/a	n/a	n/a	[B.29,30]
			1191.795	±25	?	TD-AltBOC(15,10) ^j						[B.29–31]
	B2b-D ^l		1207.14	±10	I		10.23	?	?	50/100	?	[B.29–31]
		B2b-P ^l		±10	Q		10.23	?	?	–	?	[B.29–31]
		B2a-D ^l	1176.45	±10	I		10.23	?	?	25/50	?	[B.29–31]
		B2a-P ^l		±10	Q		10.23	?	?	–	?	[B.29–31]

Abbreviations: Sys = System; BW = bandwidth; Ch = channel; LFSR = linear feedback shift register; n/a = nonavailability of public information for regulated/military services;

OS = open service; CS = commercial service; PRS = public regulated service; RS = restricted service

Notes: ⁱ combined signal; ^j GEO; ^k MEO/IGSO; ^l open service (planned); ^m authorized service (planned)

Table B.3 (continued)

Sys	Band	Signal	Frequency (MHz)	BW (MHz)	Ch	Modulation	Rate (MHz)	Code prim./second. (chips)	Type	Data (bps/sps)	Power (min. rcvdl.) (dBW)	References
QZSS-1	L1	C/A	1575.42	±1	I ^{±n}	BPSK(1)	1.023	1023	Gold	50/50	-158.5	[B.32, 33]
					Q ^{±o}							[B.34]
		L1C-D		±2	I [±]	BOC(1,1) ⁿ	1.023	10 230	Weil	50/100	-163.0	[B.32, 33]
						TMBOC(6,1,4/33) ^o						[B.34]
		L1C-P		±2	Q ^{±n}	BOC(1,1) ⁿ	1.023	10 230/1800	Weil	-	-158.2	[B.32, 33]
					I ^{±o}	TMBOC(6,1,4/33) ^o						[B.34]
		SAIF		±1	-	BPSK(1)	1.023	1023	Gold	500/250	-161.0	[B.32]
	L2	L2 CM	1227.60	±1	-	BPSK(1) mux	0.5115	10 230	M-seq.	25/50	-163.0	[B.32, 34]
		L2 CL		±1	-	BPSK(1) mux	0.5115	767 250	M-seq.	-	-163.0	[B.32, 34]
	L6	L61(LEX) ⁿ	1278.75	±5	-	BPSK(5) mux	5.115	10 230	Kasami	2000/250	-158.7	[B.32, 35]
				±5	-	BPSK(5) mux	5.115	1 048 575	Kasami	-	-158.7	[B.32, 35]
		L62 ^o		±5	-	BPSK(5) mux	5.115	10 230	Kasami	2000/250	-159.8	[B.35]
				±5	-	BPSK(5) mux	5.115	10 230	Kasami	2000/250	-159.8	[B.35]
	L5	L5I	1176.45	±10	I [±]	BPSK(10)	10.23	10 230/10	M-seq.	50/100	-157.9 ⁿ , -157.0 ^o	[B.32, 34]
		L5Q		±10	Q [±]	BPSK(10)	10.23	10 230/20	M-seq.	-	-157.9 ⁿ , -157.0 ^o	[B.32, 34]
IRNSS	L5	SPS	1176.45	±24	I ⁻	BPSK(1)	1.023	1023	M-seq.	25/50	-159.0	[B.36, 37]
		RS-D		±16	Q ⁻	BOC(5,2)	2.046	8192	n/a	25/50	-156.0	[B.36, 37]
		RS-P		±16	I ⁻	BOC(5,2)	2.046	8192/40	n/a	-	-159.0	[B.36, 37]
	S	SPS	2492.028	±16	I ⁻	BPSK(1)	1.023	1023	M-seq.	25/50	-162.3	[B.36, 37]
		RS-D		±16	Q ⁻	BOC(5,2)	2.046	8192	n/a	25/50	-159.3	[B.36, 37]
		RS-P		±16	I ⁻	BOC(5,2)	2.046	8192/40	n/a	-	-162.3	[B.36, 37]
SBAS	L1	C/A	1575.42	±1	I	BPSK(1)	1.023	1023	Gold	250/500	-161.0 ^p	[B.11, 38]
		C/A ^q		±1	Q	BPSK(1)	1.023	1023	Gold	≤250/500	-161.0 ^p	[B.11, 39]
	L5	L5I	1176.45	±10	I	BPSK(10)	10.23	10 230/2	M-seq.	250/500	-161.0 ^p	[B.11, 39]
		L5Q ^q		±10	Q	BPSK(10)	10.23	10 230/?	M-seq.	≤250/500 ^r	-161.0 ^p	[B.11, 39]

Abbreviations: Sys = System; BW = bandwidth; Ch = channel; mux = multiplexed; n/a = nonavailability of public information for regulated/military services; RS = restricted service; SPS = standard positioning service (open)
Notes: ⁿ Block I; ^o Block II; ^p Specified minimum received power of all signals; ^q Optional signal component; ^r Secondary code length varies with selected data rate, product is fixed at 500 bps

References

- B.1 IAU (1991) Recommendation IV, The Terrestrial Time (TT) (XXIst General Assembly of the International Astronomical Union, Buenos Aires, 1991) <http://www.iers.org/IIERS/EN/Science/Recommendations/recommendation4.html>
- B.2 Navstar GPS Space Segment/Navigation User Interfaces, Interface Specification IS-GPS-200, 24 Sep. 2013 (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo, California, 2013)
- B.3 BeiDou Navigation Satellite System Signal In Space Interface Control Document – Open Service Signal, v2.1, Nov. 2016 (China Satellite Navigation Office, Beijing 2016)
- B.4 P.J. Mohr, D.B. Newell, B.N. Taylor: CODATA Recommended Values of the Fundamental Physical Constants: 2014 (National Institute of Standards and Technology, Gaithersburg 2015) <http://arxiv.org/abs/1507.07956>
- B.5 N.K. Pavlis, S.A. Holmes, S.C. Kenyon, J.K. Factor: The development and evaluation of the Earth Gravitational Model 2008 (EGM2008), J. Geophys. Res. (2012), doi:10.1029/2011JB008916
- B.6 H. Moritz: Geodetic Reference System 1980, J. Geod. **74**(1), 128–133 (2000)
- B.7 W.M. Folkner, J.G. Williams, D.H. Boggs: The Planetary and Lunar Ephemeris DE 421, IPN Progress Report 42–178 (Jet Propulsion Laboratory, Pasadena 2009)
- B.8 E.V. Pitjeva, E.M. Standish: Proposals for the masses of the three largest asteroids, the Moon–Earth mass ratio and the Astronomical Unit, Celest. Mech. Dyn. Astron. **103**(4), 365–372 (2009)
- B.9 B.A. Archinal, M.F. A'Hearn, E. Bowell, A. Conrad, G.J. Consolmagno, R. Courtin, T. Fukushima, D. Hestroffer, J.L. Hilton, G.A. Krasinsky, G. Neumann, J. Oberst, P.K. Seidelmann, P. Stooke, D.J. Tholen, P.C. Thomas, I.P. Williams: Report of the IAU Working Group on cartographic coordinates and rotational elements: 2009, Celest. Mech. Dyn. Astron. **109**(2), 101–135 (2010)
- B.10 G. Kopp, J.L. Lean: A new, lower value of total solar irradiance: Evidence and climate significance, Geophys. Res. Lett. (2011), doi:10.1029/2010GL045777
- B.11 J. Betz: *Engineering Satellite-Based Navigation and Timing – Global Navigation Satellite Systems, Signals, and Receivers* (Wiley-IEEE, Hoboken 2016)
- B.12 P. Ward: GPS satellite signal characteristics. In: *Understanding GPS – Principles and Applications*, 2nd edn., ed. by E.D. Kaplan, C.J. Hegarty (Artech House, Norwood 2006) pp. 83–117, Chap. 4
- B.13 Navstar GPS Space Segment/User Segment L1C Interfaces, Interface Specification IS-GPS-800D, 24 Sep. 2013 (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo, California, 2013)
- B.14 B.C. Barker, J.W. Betz, J.E. Clark, J.T. Correia, J.T. Gillis, S. Lazar, K.A. Rehborn, J.R. Straton III: Overview of the GPS M code signal, Proc. ION NTM 2000, Anaheim (ION, Virginia 2000) pp. 542–549
- B.15 W.A. Marquis, D.L. Reigh: The GPS Block IIR and IIR-M broadcast L-band antenna panel: Its pattern and performance, Navigation **62**(4), 329–347 (2015)
- B.16 Navstar GPS Space Segment/User Segment L5 Interfaces, Interface Specification IS-GPS-705D, 24 Sep. 2013 (Global Positioning Systems Directorate, Los Angeles Air Force Base, El Segundo, California, 2013)
- B.17 B.A. Stein, W.L. Tsang: PRN codes for GPS/GLONASS: A Comparison, ION NTM 1990, San Diego (ION, Virginia 1990) pp. 31–35
- B.18 J. Beser, J. Danaher: The 3S Navigation R-100 Family of Integrated GPS/GLONASS Receivers: Description and Performance Results, ION NTM 1993, San Francisco (ION, Virginia 1993) pp. 25–45
- B.19 Global Navigation Satellite System GLONASS – Interface Control Document, v5.1 (Russian Institute of Space Device Engineering, Moscow 2008)
- B.20 Y. Urlichich, V. Subbotin, G. Stupak, V. Dvorkin, A. Povaliaev, S. Karutin: GLONASS developing strategy, ION GNSS 2010, Portland (ION, Virginia 2010) pp. 1566–1571
- B.21 Y. Urlichich, V. Subbotin, G. Stupak, V. Dvorkin, A. Povaliaev, S. Karutin: GLONASS modernization, ION GNSS 2011, Portland (ION, Virginia 2011) pp. 3125–3128
- B.22 Global Navigation Satellite System GLONASS Interface Control Document, L1 Open Access Code Division Radio Navigation Signal, v1.0 (JSC Russian Space Systems, Moscow 2016) in Russian
- B.23 Global Navigation Satellite System GLONASS Interface Control Document, L2 Open Access Code Division Radio Navigation Signal, v1.0 (JSC Russian Space Systems, Moscow 2016) in Russian
- B.24 S. Revnivkykh: GLONASS status and evolution, IAIN World Congress, Prague (IAIN, Netherlands 2015) <http://www.iainav.org/iaain-iwc2015/iaain-2015-keynote-lecture-revnivkykh.pdf>
- B.25 European GNSS (Galileo) Open Service Signal In Space Interface Control Document, OS SIS ICD, Iss. 1.3, Dec. 2016 (EU, Brussels 2016)
- B.26 J.-A. Avila-Rodriguez, G.W. Hein, S. Wallner, J.-L. Issler, L. Ries, L. Lestarquit, A. de Latour, J. Godet, F. Bastide, T. Pratt, J. Owen: The MBOC modulation: The final touch to the Galileo frequency and signal plan, Navigation **55**(1), 15–28 (2008)
- B.27 T. Grelier, J. Dantepal, A. Delatour, A. Ghion, L. Ries: Initial observations and analysis of Compass MEO satellite signals, Inside GNSS **2**(4), 39–43 (2007)
- B.28 G.X. Gao, A. Chen, S. Lo, D. De Lorenzo, T. Walter, P. Enge: Compass-M1 broadcast codes in E2, E5b, and E6 frequency bands, IEEE J. Sel. Top. Signal Process. **3**(4), 599–612 (2009)
- B.29 Description of systems and networks in the radionavigation-satellite service (space-to-Earth and space-to-space) and technical characteristics of transmitting space stations operating in the bands 1164–1215 MHz, 1215–1300 MHz and 1559–1610 MHz, Recommendation M 1787, rev. 2, Sep. 2014 (ITU, Geneva 2014) <https://www.itu.int/rec/R-REC-M.1787/en>
- B.30 W. Xiao, W. Liu, G. Sun: Modernization milestone: BeiDou M2-S initial signal analysis, GPS Solutions **20**(1), 125–133 (2016)

- B.31 Z.P. Tang, H.W. Zhou, J.L. Wei, T. Yan, Y.Q. Liu, Y.H. Ran, Y.L. Zhou: TD-AltBOC: A new COMPASS B2 modulation, *Sci. China Phys. Mech. Astron.* **54**(6), 1014–1021 (2011)
- B.32 Quasi-Zenith Satellite System Navigation Service Interface Specification for QZSS, IS-QZSS, v1.6, 28 Nov. 2014 (JAXA, 2014)
- B.33 H. Maeda: System Research on The Quasi-Zenith Satellites System, Ph.D. Thesis (Tokyo University of Marine Science and Technology, Tokyo 2007), in Japanese
- B.34 Quasi-Zenith Satellite System Interface Specification – Satellite Positioning, Navigation and Timing Service, IS-QZSS-PNT-001, Draft 12 July 2016 (Cabinet Office, Tokyo 2016)
- B.35 Quasi-Zenith Satellite System Interface Specification – Centimeter Level Augmentation Service, IS-QZSS-L6-001, Draft 12 July 2016 (Cabinet Office, Tokyo 2016)
- B.36 Indian Regional Navigation Satellite System – Signal In Space ICD for Standard Positioning Service, version 1.0, June 2014 (Indian Space Research Organization, Bangalore 2014)
- B.37 S. Thielert, O. Montenbruck, M. Meurer: IRNSS-1A – Signal and clock characterization of the Indian Regional Navigation System, *GPS Solutions* **18**(1), 147–152 (2014)
- B.38 Minimum Operational Performance Standards for GPS/WAAS Airborne Equipment, RTCA DO-229D, 13 Dec. 2006 (RTCA, Washington, DC 2006)
- B.39 Signal Specification for SBAS L1/L5, ED-134, Draft v.3, May 2008 (The European Organisation for Civil Aviation Equipment, Paris 2008)

About the Authors



Zuheir Altamimi

Chapter F.36

Institut National de l'Information
Géographique et Forestière (IGN)
Université Paris Diderot (LAREG)
Paris, France
zuheir.altamimi@ign.fr

Zuheir Altamimi is Research Director at the Institut National de l'Information Géographique et Forestière (IGN), France. His research focuses are space geodesy and theory and realization of terrestrial reference systems. He is Head of the IGN Terrestrial Reference Systems Research group and the International Terrestrial System (ITRS) Center. He received his PhD in Space Geodesy from Paris Observatory and his habilitation from Paris University VI.

Felix Antreich

Federal University of Ceará (UFC)
Dept. of Teleinformatics Engineering
Fortaleza, Brazil
antreich@ieee.org



Chapter A.4

Felix Antreich received a Doktor-Ingenieur (PhD) in Electrical Engineering from Munich University of Technology (TUM), Germany, in 2011. Since July 2003, he has been an Associate Researcher with the Department of Navigation, Institute of Communications and Navigation of the German Aerospace Center (DLR), Oberpfaffenhofen, Germany. His research interests include sensor array signal processing for global navigation satellite systems (GNSS) and wireless communications, estimation theory, and signal design for synchronization and GNSS.

Ron Beard

US Naval Research Laboratory
Advanced Space PNT Branch
Washington, USA
ronald.beard@verizon.net



Chapter A.5

Ronald Beard is the former Head of the NRL Advanced Space PNT Branch. In the 1970s, he was the project scientist in the NRL GPS Program Office that developed Navigation Technology Satellites 1 and 2, which operated the first atomic clocks in space. He became the Program Manager of the NRL GPS Clock Development program in 1984 and Precise Time and Time Interval (PTTI) technology for military navigation and communication systems.

Alexey Bolkunov

Federal Space Agency (Roscosmos)
PNT Information and Analysis Center
Korolyov, Russian Federation
alexei.bolkunov@glonass-iac.ru



Chapter B.8

Alexey Bolkunov is a Senior Research Associate of the PNT Information and Analysis Center of the Central Scientific Research Institute for Machine Building, Federal Space Agency (Roscosmos), Russia, which he joined as a graduate of Moscow Aviation Institute (National Research University) in 2007. He received his PhD in 2011 from the same institute. Alexey Bolkunov's research interests include GNSS performance and PNT legal and regulatory framework studies.



Michael S. Braasch

Chapter C.15

Ohio University
School of Electrical Engineering &
Computer Science
Athens, USA
braaschm@ohio.edu

Michael S. Braasch holds the Thomas Professorship in the Ohio University School of Electrical Engineering and Computer Science and is Principal Investigator with the Ohio University Avionics Engineering Center (AEC). One focus of his research has been the characterization and mitigation of the effects of multipath in GNSS. Other areas of research include GNSS software-defined receiver development, novel general aviation cockpit displays and unmanned aerial vehicle sense-and-avoid systems.

Thomas Burger

European Space Agency (ESA)
Galileo Project Office
Noordwijk, The Netherlands
thomas.burger@esa.int



Chapter B.9

Thomas Burger is Principal Signal Engineer within the Galileo Project Office at the European Space Agency in Noordwijk (The Netherlands). He received his PhD in Engineering Sciences from the Technical University of Darmstadt, Germany. He has been working on space systems, focussed on space-based navigation applications and systems, synthetic aperture radar and microwave instruments, and signal processing for precision applications.

Estel Cardellach

Institute of Space Sciences
ICE (IEEC-CSIC)
Cerdanyola del Valles, Spain
estel@ice.csic.es



Chapter G.40

Estel Cardellach received a PhD from UPC, Barcelona, Spain (2002). She works on scientific applications of the Global Navigation Satellite Systems (GNSS) for remote sensing of the Earth, including reflectometry and radio-occultations. She enjoyed postdoctoral positions at Jet Propulsion Laboratory and Harvard Smithsonian Center for Astrophysics. Currently she works at the Institute of Space Sciences (ICE-CSIC/IEEC), Spain.

James T. Curran



European Space Agency (ESA)
Noordwijk, The Netherlands
jamestcurran@ieee.org

Chapter C.18

James T. Curran received a Bachelor degree in Electrical Engineering and a Doctorate in Telecommunications, from University College Cork, Ireland. Since then he has worked as a researcher at various institutions, including the PLAN Group at the University of Calgary and at the Joint Research Centre of the European Commission, Italy. Although focusing on radio-navigation, his research interests also include signal processing, information theory, cryptography and software defined radio.

Pascale Defraigne



Royal Observatory of Belgium
Brussels, Belgium
p.defraigne@oma.be

Chapter G.41

Pascale Defraigne received her PhD in Physics from the Université Catholique de Louvain (UCL), Belgium (1995). She was Assistant at the Royal Observatory of Belgium in the field of GNSS Time and Frequency transfer and is now Head of the Time Laboratory. She participated in the development of the European Navigation System Galileo and chairs the working group on GNSS Time Transfer of the Consultative Committee of Time and Frequency.

Bernd Eissfeller

Universität der Bundeswehr München
Inst. of Space Technology and Space
Applications
Neubiberg, Germany
bernd.eissfeller@unibw.de



Chapter C.13

Bernd Eissfeller received his PhD in Inertial Geodesy (1989) and his Habilitation (venia legendi) in Navigation and Physical Geodesy (1996). He is Full Professor of Navigation at the Institute of Space Technology and Space Applications at the University of the Bundeswehr (Federal Armed Forces/UFAF), Munich. His current research interests are in the field of Galileo system optimization, GNSS receiver design, GNSS/INS multi-sensor hybridization and deep-space geodesy.

Gunnar Elgered

Chalmers University of Technology
Dept. of Earth and Space Sciences
Onsala, Sweden
gunnar.elgered@chalmers.se



Chapter G.38

Gunnar Elgered received the MSEE (1977) and PhD (1983) degrees from Chalmers University of Technology, Gothenburg, Sweden. He is Professor of Electrical Measurements and Head of the department. His research area is space geodesy, using radio telescopes and global navigational satellite systems (GNSS). More recently, his focus has been on the use of GNSS data for climate research and evaluation of climate models.

Marco Falcone



European Space Agency (ESA)
Galileo Project Office
Noordwijk, The Netherlands
marco.falcone@esa.int

Chapter B.9

Marco Falcone is the System Manager in the Galileo Project Office at the European Space Agency in Noordwijk (The Netherlands). He received his Master's in Computer Science from the University of Pisa, Italy (1987) and a Master's in Space Systems Engineering from the University of Delft, The Netherlands (1999). He has 28 years' work experience in system engineering mainly applied to large space systems.

**Richard Farnworth**

Chapter E.30

Eurocontrol Experimental Centre
Centre du Bois des Bordes – BP15
Bretigny sur Orge, France
richard.farnworth@eurocontrol.int

Richard Farnworth is the Deputy Head of the Navigation and CNS Research Unit within the Eurocontrol Directorate of Air Traffic Management. He supports the deployment of performance-based navigation and contributes to several projects addressing approaches with vertical guidance. He began his career with Eurocontrol at its Experimental Centre in 1996, where he was employed as an expert in satellite navigation looking at how to use GNSS in aviation.

Jay A. Farrell

Chapter E.28

University of California, Riverside
Dept. of Electrical and Computer
Engineering
Riverside, USA
farrell@ece.ucr.edu



Jay A. Farrell earned BS degrees from Iowa State University, and MS and PhD degrees in Electrical Engineering from the University of Notre Dame. He has worked at Draper Lab and is a Professor and two time Chair of the Department of Electrical and Computer Engineering at the University of California, Riverside. His research interests relate to control, state estimation, sensor fusion, and planning for autonomous vehicle applications.

Jeff Freymueller

Chapter F.37

University of Alaska
Geophysical Institute
Fairbanks, USA
jfreymueller@alaska.edu



Jeff Freymueller received his PhD in Geology in 1991 from the University of South Carolina. After a postdoctoral fellowship at Stanford University, he has been on the faculty at the University of Alaska Fairbanks since 1995. Throughout his career he has worked on the application of space geodesy to the study of deformation of the Earth, including studies in tectonics, earthquakes and the earthquake cycle, volcanism, and changes in cryospheric loads.

A.S. Ganeshan

Chapter B.11



Indian Space Research Organization (ISRO)
ISRO Satellite Centre (ISAC)
Bangalore, India
asganesan53@gmail.com

A.S. Ganeshan is the former Programme Director of Satellite Navigation Program and Group Director, Space Navigation Group at the ISRO Satellite Centre, Bangalore. He also served as Executive Head of the realization of the certified GAGAN system over India. He originated the concept of regional navigation satellite system (IRNSS) and has played a key role in the realization of the same. He is currently associated with Airports Authority of India as Advisor on GAGAN related matters.

Steven Gao

Chapter C.17



University of Kent
School of Engineering and Digital Arts
Canterbury, Kent, UK

Steven Gao is Professor and Chair of RF and Microwave Engineering at the University of Kent, UK. His research covers satellite antennas, smart antennas, phased arrays, GNSS antennas, RF/microwave/millimetre-wave/THz circuits, satellite communication, radars (synthetic aperture radar, UWB radars) and small satellites. He has published over 200 papers and 2 books.

Gabriele Giorgi

Chapter E.27

Technical University of Munich
Inst. for Communications and Navigation
Munich, Germany
gabriele.giorgi@tum.de



Gabriele Giorgi is a lecturer and researcher at the Institute for Communications and Navigation, Technical University of Munich, Germany. He obtained a PhD following his work on global navigation satellite system (GNSS) for aerospace applications at the Delft Institute of Earth Observation and Space Systems (DEOS), Delft University of Technology (The Netherlands). His current research focuses on reliable GNSS positioning for aeronautics applications.

Richard Gross

California Institute of Technology
Jet Propulsion Laboratory
Pasadena, USA
richard.s.gross@jpl.nasa.gov



Chapter F.36

Richard Gross has over 30 years' experience in space geodesy. His research interests include Earth rotation, time variable gravity and, most recently, terrestrial reference frame determination. He has published over 60 peer-reviewed articles on these topics and has worked at the Jet Propulsion Laboratory, California Institute of Technology since 1988, where he is a Senior Research Scientist and, since 2006, the Supervisor of the Geodynamics and Space Geodesy Group.

Jörg Hahn



European Space Agency (ESA)
Galileo Project Office
Noordwijk, The Netherlands
joerg.hahn@esa.int

Chapter B.9

Jörg Hahn is Head of the Galileo System Procurement Service in the Galileo Project Office at the European Space Agency in Noordwijk. He received his PhD in Engineering Sciences from the University of Federal Armed Forces Munich-Neubiberg, Germany (1999). He has 22 years' work experience in system engineering, mainly with timing, satellite navigation, and large space systems.

André Hauschild



German Aerospace Center (DLR)
German Space Operations Center
Wessling, Germany
andre.hauschild@dlr.de

Chapters D.19, D.20

André Hauschild is a researcher in the GNSS Technology and Navigation Group at DLR's German Space Operations Center (GSOC). He received his PhD from the Technical University Munich, Germany, in 2010. His field of work focuses on real-time clock offset estimation for GNSS satellites for precise positioning, multi-GNSS processing using modernized and new satellite navigation systems, as well as space-borne GNSS applications.

Grant Hausler

Geoscience Australia
Symonston, Australia
grant.hausler@ga.gov.au



Chapter F.33

Grant Hausler is Coordinator for the National Positioning Infrastructure at Geoscience Australia. He holds a Bachelor of Geomatic Engineering and a PhD from the University of Melbourne. Grant provides Secretariat to the Australian Government Positioning, Navigation and Timing Working Group and the National Positioning Infrastructure Advisory Board, and represents Geoscience Australia on the Attorney General's Space Community of Interest.

Christopher J. Hegarty

The MITRE Corporation
Bedford, USA
chegarty@mitre.org



Chapter B.7

Christopher J. Hegarty is the Director for CNS Engineering & Spectrum with The MITRE Corporation, where he has worked mainly on aviation applications of GNSS since 1992. He received BS and MS degrees in Electrical Engineering from WPI and a DSc degree in EE from GWU. He is currently the Chair of the Program Management Committee of RTCA, Inc., and co-chairs RTCA Special Committee 159 (GNSS).

Thomas Hobiger



Chalmers University of Technology
Onsala Space Observatory
Onsala, Sweden
thomas.hobiger@chalmers.se

Chapter A.6

Thomas Hobiger is Associate Professor for Space Geodesy. He received the MSc and PhD degrees in Geodesy and Geophysics from the Vienna University of Technology, Austria (2002, 2005). He spent 8 years at a Japanese research institute, before moving to Chalmers in 2014. He plays an active role in the development of next-generation space-geodetic systems (in particular VLBI and GNSS) and processing tools for such techniques.



Urs Hugentobler

Technical University of Munich
Satellite Geodesy
Munich, Germany
urs.hugentobler@bv.tum.de

Chapter A.3

Urs Hugentobler has been Professor for Satellite Geodesy at the Technical University Munich, Germany and Head of the Research Facility Satellite Geodesy since 2006. He obtained his PhD in Astronomy from the University of Bern, Switzerland, in 1998. His research activities include precise positioning applications using GNSS, precise orbit determination and modeling, and clock modeling, using the new GNSS satellite systems.

Todd Humphreys

The University of Texas at Austin
Aerospace Engineering and Engineering
Mechanics,
W.R. Woolrich Laboratories, C0600
Austin, USA
todd.humphreys@mail.utexas.edu



Chapter C.16

Todd Humphreys (BS, MS, Electrical Engineering, Utah State University; PhD, Aerospace Engineering, Cornell University) is Associate Professor in the Department of Aerospace Engineering and Engineering Mechanics at the University of Texas at Austin, where he directs the UT Radionavigation Laboratory. He specializes in the application of optimal detection and estimation techniques to problems in satellite navigation, autonomous systems, and signal processing.

Norbert Jakowski

German Aerospace Center (DLR)
Institute of Communications and
Navigation
Neustrelitz, Germany
norbert.jakowski@dlr.de



Chapters A.6, G.39

Norbert Jakowski received his PhD in Solid State Physics in 1974 from the University of Rostock. Since 1974 he has been working in the Institute of Space Research and since 1991 at the German Aerospace Center (DLR) at their branch in Neustrelitz. His research activities include monitoring, modeling and predicting ionospheric processes related to space weather events and studying their impact on radio wave propagation, in particular on GNSS applications.



Christopher Jekeli

Ohio State University
School of Earth Sciences
Columbus, USA
jekeli.1@osu.edu

Chapter A.2

Christopher Jekeli received his PhD degree in Geodetic Science from the Ohio State University in 1981. He was employed by the U.S. Air Force Geophysical Laboratory as geodesist (1981–1993), joined the faculty of the Department of Geodetic Science, and the School of Earth Sciences, at the Ohio State University (2005). His principal research interests are geodesy and the Earth's gravity field, its modeling and measurement for geodetic and geophysical applications.



Gary Johnston

Geoscience Australia
Symonston, Australia
gary.johnston@ga.gov.au

Chapter F.33

Gary Johnston leads the Geodesy and Seismic Monitoring Group at Geoscience Australia. He is the Chair of the International GNSS Service (IGS) Governing Board and the Chair of the UN Global Geospatial Information Management (UN GGIM) working group on the Global Geodetic Reference Frame (GGRF).

Allison Kealy

University of Melbourne
Dept. of Infrastructure Engineering
Parkville, Australia
akealy@unimelb.edu.au



Chapter E.29

Dr Allison Kealy is an Associate Professor in the Department of Infrastructure Engineering at The University of Melbourne Australia. She holds a PhD in GPS and Geodesy from the University of Newcastle upon Tyne, UK. Allison's research interests include sensor fusion, Kalman filtering, high precision satellite positioning, GNSS quality control, wireless sensor networks and location based services.

Satoshi Kogure

National Space Policy Secretariat,
Cabinet Office
QZSS Strategy Office
Tokyo, Japan
satoshi.kogure.e7f@cao.go.jp



Chapter B.11

Satoshi Kogure is the former Mission Manager for QZSS in the Satellite Navigation Unit, JAXA. He received an MS from Nagoya University in 1993 and joined the Japanese National Space Agency. He started system design work for the Japanese navigation satellite system QZSS in 2001 and led its technical validation and demonstration after the first satellite launch in 2010. Since 2016, he coordinates the QZS programme within the Japanese Cabinet Office.



Jan Kouba

Natural Resources Canada
Canadian Geodetic Survey
Ottawa, Canada
kouba@rogers.com

Chapter E.25

Jan Kouba obtained his Doctorate of Science from the Czech Academy of Sciences in 1994. He has been working in satellite geodesy since 1970. He has held several research positions at the Geological Survey of Canada and the Geodetic Survey Division (GSD) of the Natural Resources Canada (NRCan). During 1994–1998 he led the Canadian Active Control Technology/IGS (International GPS Service) Analysis Team at GSD and served as the first Analysis Center Coordinator of IGS.



François Lahaye

Natural Resources Canada
Canadian Geodetic Survey
Ottawa, Canada
francois.lahaye@canada.ca

Chapter E.25

François Lahaye leads the Geodetic Space-based Technology team at the Canadian Geodetic Survey of Natural Resources Canada. He holds an MSc in geodetic sciences from Laval University, Canada.

Richard B. Langley

University of New Brunswick
Dept. of Geodesy & Geomatics
Engineering
Fredericton, Canada
lang@unb.ca



Chapter A.1

Richard B. Langley is Professor in the Department of Geodesy and Geomatics Engineering at the University of New Brunswick in Fredericton, Canada, where he has been teaching and conducting research since 1981. He holds a PhD in Experimental Space Science from York University, Toronto. His general area of expertise is precision applications of GNSS and he has been active in the development of GNSS error models since the early 1980s.

Ken MacLeod

Natural Resources Canada
Canadian Geodetic Survey
Ottawa, Canada
ken.macleod@canada.ca



Annex A

Ken MacLeod received a Bachelor of Science degree from the University of Toronto in 1985. He joined the Canadian Geodetic survey in 1987 and since 1995 he has worked on the development of GNSS augmentation systems. He is currently the IGS/RTCM SC104 RINEX working group chairman.



Moazam Maqsood

Institute of Space Technology
Dept. of Electrical Engineering
Islamabad, Pakistan
moazam.maqsood@ist.edu.pk

Chapter C.17

Moazam Maqsood received his Communication Systems Engineering degree from the Institute of Space Technology (IST), Islamabad, Pakistan in 2006, and MS and PhD degrees from the University of Surrey, UK, in 2009 and 2013. He is currently serving as Assistant Professor in the Electrical Engineering Department at IST. His research interests include use of synthetic aperture radars (SAR) for public safety applications.



Michael Meurer

German Aerospace Center (DLR)
Institute of Communications and
Navigation
Wessling, Germany
michael.meurer@dlr.de

Chapter A.4; Annex B

Dr Michael Meurer is Head of the Department of Navigation of the German Aerospace Center (DLR), Institute of Communications and Navigation, and the Coordinating Director of the DLR Center of Excellence for Satellite Navigation. He is also Professor of Electrical Engineering and Director of the Chair of Navigation at RWTH Aachen. His current research interests include GNSS signals, GNSS receivers, interference and spoofing mitigation, and navigation for safety-critical applications.

Oliver Montenbruck Chaps. A.1, A.2, A.3, B.8, B.10, B.11, C.17, E.32; Annexes A, B For biographical profile, please see the section “About the Editors”.

Terry Moore

University of Nottingham
Nottingham Geospatial Institute
Nottingham, UK
terry.moore@nottingham.ac.uk

**Chapter E.29**

Professor Terry Moore is Director of the Nottingham Geospatial Institute (NGI) at the University of Nottingham. He holds a BSc degree in Civil Engineering and a PhD degree in Space Geodesy, both from the University of Nottingham. He is a Fellow, and a Member of Council, of both the Institute of Navigation and of the Royal Institute of Navigation.

Dennis Odijk

Fugro Intersite B.V.
Leidschendam, The Netherlands
d.odijk@fugro.com

**Chapters D.21, E.26**

Dennis Odijk received his PhD degree in Geodetic Engineering from Delft University of Technology, the Netherlands (2002). From 2009 to 2016 he was a Research Fellow at the GNSS Research Centre at Curtin University in Perth, Australia, where he focused on RTK and integer ambiguity resolution enabled PPP. At present Dennis is working as Senior Geodesist with Fugro Intersite in the Netherlands, where he contributes to Fugro's GNSS positioning algorithms.

**Thomas Pany**

Universität der Bundeswehr München
Inst. of Space Technology and Space
Applications
Neubiberg, Germany
thomas.pany@unibw.de

Chapter C.14

Prof. Thomas Pany is with the Universität der Bundeswehr München where teaches navigation. He obtained a PhD from the Graz University of Technology. His research focuses on GNSS signals, GNSS receiver design and integration with other sensors. Previously he worked for IFEN GmbH and is the architect of the world's most powerful GNSS software receiver.

**Mark G. Petovello**

University of Calgary
Geomatics Engineering
Calgary, Canada
mark.petovello@ucalgary.ca

Chapter C.18

Mark G. Petovello received his BSc and PhD degrees in Geomatics Engineering at the University of Calgary in 1998 and 2003, respectively. He is currently a Professor at the same university with research interests in satellite-based navigation, inertial navigation, and multi-sensor integration. He has written and licensed several navigation-related software packages and is actively involved in the navigation community.

Sam Pullen

Stanford University
Dept. of Aeronautics and Astronautics
Stanford, USA
spullen@stanford.edu

**Chapter E.31**

Sam Pullen is Technical Manager of the Ground Based Augmentation System (GBAS) research effort within the GNSS Laboratory at Stanford University, where he received his PhD in Aeronautics and Astronautics in 1996. He has supported FAA and other service providers in developing system concepts, technical requirements, integrity algorithms, and performance models for GBAS, SBAS, and other GNSS applications. He has performed GNSS system design, applications, and risk assessment through his consultancy.

Sergey Revnivkykh

RESHETNEV's Information Satellite
Systems Coporation
GLONASS Evolution Department
Moscow, Russian Federation
revnivkykh@iss-reshetnev.ru

**Chapter B.8**

Sergey Revnivkykh is Deputy Head of the GLONASS Directorate and Director of the GLONASS Evolution Department of Reshetnev's Information Satellite Systems Corporation. Since 2001 he has been actively involved in the GLONASS Federal Program development, management, and implementation. Since 2005 he has been a co-chair of the Working Group on GNSS Compatibility and Interoperability of the International Committee on GNSS. He received his PhD from Moscow Aviation Institute (2006).



Anna Riddell

Chapter F.33

Geoscience Australia
Symonston, Australia
anna.riddell@ga.gov.au

As a graduate of the University of Tasmania, Anna started her career in geodesy at Geoscience Australia, where she is currently the Primary Operator of the robotic GNSS antenna calibration facility. Other research activities include GNSS analysis for the production of the Asia Pacific Reference Frame (APREF) and providing legal traceability for the verification of position within Australia.



Antonio Rius

Chapter G.40

Institute of Space Sciences
ICE (IEEC-CSIC)
Cerdanyola del Valles, Spain
rius@ice.csic.es

Antonio Rius received a PhD degree from Barcelona University, Spain, in 1974. From 1975 to 1985, he was a member of the technical staff at NASA's Deep Space Communications Complex, Madrid, Spain, where he was responsible for radio astronomical activities. Since 1986, he has been with the Spanish Research Council and the Institut d'Estudis Espacials de Catalunya.

Chris Rizos



Chapter F.35

The University of New South Wales
School of Civil & Environmental
Engineering
Kensington, Australia
c.rizos@unsw.edu.au

Chris Rizos is Professor of Geodesy and Navigation, School of Civil and Environmental Engineering, University of New South Wales, Australia. Chris is immediate Past President of the International Association of Geodesy (IAG), a member of the Governing Board of the International GNSS Service (IGS), and co-chair of the Multi-GNSS Asia Steering Committee. Chris has been researching the technology and applications of GPS and other navigation/positioning systems since 1985.

Ken Senior



Chapter A.5

US Naval Research Laboratory
Advanced Space PNT Branch
Washington, USA
ken.senior@nrl.navy.mil

Kenneth L. Senior received his PhD in Applied Mathematics from Bowling Green State University, in 1997. He joined the Naval Research Laboratory as a scientist in 2001 and currently heads the Advanced Space Positioning Navigation and Timing Branch of the Naval Research Laboratory. His research interests include precision time and frequency transfer, timescales, and the analysis of precision clocks.

Alexander Serdyukov (deceased)

Chapter B.8



Tim Springer

Chapter F.34

PosiTim UG
Seeheim-Jugenheim, Germany
tim.springer@positim.com

Tim Springer studied Aerospace Engineering at the Technical University of Delft, and worked at the Astronomical Institute of the University of Bern, where he received his PhD in Physics in 1999. He started the company PosiTim, which offers services and solutions for high accuracy GNSS marked based on the ESA/ESOC software NAPEOS. He has been working at the Navigation Support Office of ESA/ESOC in Darmstadt since 2004.



Peter Steigenberger

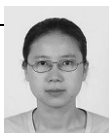
Chapter F.34; Annex B

German Aerospace Center (DLR)
German Space Operations Center
Wessling, Germany
peter.steigenberger@dlr.de

Peter Steigenberger received his Master's and PhD degrees in Geodesy from the Technical University of Munich (TUM) in 2002 and 2009, respectively. Currently, he is Scientific Staff Member at DLR's German Space Operations Center (GSOC). His research interests focus on GNSS data analysis, in particular precise orbit and clock determination of GNSS satellites and the evolving navigation systems Galileo, BeiDou, and QZSS.

Jing Tang

China National Administration of GNSS
and Applications
Beijing, China
blazingtangjing@163.com



Chapter B.10

Jing Tang is an engineer at China National Administration of GNSS and Applications (CNAGA). She received her BA in English from Henan University of Science and Technology in 1996. Since 2007, she has been closely involved in the bilateral frequency coordination between BeiDou and other GNSSs. She has been an active participant of the Internal Committee of GNSS (ICG) meetings since 2008. She has coauthored several papers about the Beidou Satellite Navigation System.

Pierre Tétreault

Natural Resources Canada
Canadian Geodetic Survey
Ottawa, Canada
pierre.tetreault@canada.ca



Chapter E.25

Pierre Tétreault is a member of the Space Based Technology team of the Canadian Geodetic Survey at Natural Resources Canada. His work is focused on GNSS based end-user applications for reference frame access. He received an MSc with specialization in geodesy from the University of Toronto in 1987.

Peter J.G. Teunissen

Chapters A.1, D.22, D.23, D.24

For biographical profile, please see the section "About the Editors".

Sandra Verhagen

Delft University of Technology
Faculty of Civil Engineering and
Geosciences
Delft, The Netherlands
a.a.verhagen@tudelft.nl

Chapter D.22

Sandra Verhagen is Assistant Professor at Delft University of Technology. She received her PhD in Geodesy from the same university. She is Theme Leader of Sensing from Space of the Delft Space Institute, and was the president of Commission 4 Positioning and Applications of the International Association of Geodesy (2007–2011). Her research interests are mathematical geodesy and positioning, with a focus on algorithm development for very precise and reliable positioning with GNSS.

Todd Walter

Chapter B.12



Stanford University, GPS Lab
Stanford, USA
twalter@stanford.edu

Todd Walter is a Senior Research Engineer in the Department of Aeronautics and Astronautics at Stanford University. He received his PhD in Applied Physics from Stanford University in 1993. His research focuses on implementing high-integrity air navigation systems. He is active in international standards bodies coordinating the use of global navigation satellite systems to implement these systems. He is a Fellow of ION and has also served as its President.

Lambert Wanninger

Technical University Dresden
Geodetic Institute
Dresden, Germany
lambert.wanninger@tu-dresden.de



Chapter E.26

Lambert Wanninger is Professor of Geodesy at the Technical University of Dresden (TU Dresden). He holds a Dr.-Ing. degree in Geodesy from the University of Hannover, Germany, and a Habilitation degree in Geodesy from TU Dresden. He has been involved in research on precise GNSS positioning since 1990.

Jan P. Weiss

University Corporation for Atmospheric
Research
COSMIC Program
Boulder, USA
weissj@ucar.edu



Chapter F.34

Jan P. Weiss is Manager of the COSMIC Data Analysis and Archive Center at the University Corporation for Atmospheric Research in Boulder, CO. His research focuses on precise orbit and clock determination, and geodetic applications of GNSS. At JPL, he worked as an analyst and developer for the JPL IGS Analysis Center, GIPSY-OASIS software, NASA Global Differential GNSS System, and next generation GPS control segment navigation software.



Jan Wendel

Chapter E.28

Airbus DS GmbH
Navigation and Apps. Programmes
Taufkirchen, Germany
jan.wendel@airbus.com

Jan Wendel received the Dipl.-Ing. and Dr.-Ing. degrees in Electrical Engineering from the University of Karlsruhe in 1998 and 2003, respectively. From 2003 until 2006 he was Assistant Professor at the University of Karlsruhe, where he is currently a private lecturer. In 2006, he joined MBDA in Munich. In 2009, he joined EADS Astrium GmbH in Munich, Germany, now Airbus DS GmbH, where he is involved in various activities related to satellite navigation.



Jens Wickert

Chapter G.38

GFZ German Research Centre for
Geosciences
Dept. of Geodesy
Potsdam, Germany
wickert@gfz-potsdam.de

Jens Wickert graduated in physics from TU Dresden and received his PhD from Karl-Franzens-University Graz, in 2002. Since 1999 he has been with the German Research Centre for Geosciences (GFZ) and is responsible for GNSS remote sensing research. He was Principal Investigator of the pioneering GPS Radio Occultation experiment aboard the German CHAMP satellite and holds a joint professorship on GNSS Remote Sensing, Navigation and Positioning of GFZ and Technische Universität Berlin since 2016.

Jong-Hoon Won

Chapters C.13, C.14

Inha University
Faculty of Electrical Engineering
Incheon, Korea
jh.won@inha.ac.kr



Jong-Hoon Won studied control engineering at Ajou University, Suwon, South Korea (PhD). He joined the Institute of Space Technology and Space Applications (formerly the Institute of Geodesy and Navigation), University of Federal Armed Forces (UFAF) Munich, Germany in 2005. Currently, he is Assistant Professor of Electrical Engineering at Inha University, South Korea. His current research interests are in the field of GNSS signals, receivers, navigation, and target tracking systems.

Yuanxi Yang

Chapter B.10

China National Administration of GNSS
and Applications
Beijing, China
yuanxi_yang@163.com



Yuanxi Yang is Professor of Geodesy and Navigation at both Xian Research Institute of Surveying and Mapping and China National Administration of GNSS and Applications (CNAGA). He received his PhD in Geodesy from the Institute of Geodesy and Geophysics of the Chinese Academy of Science. His main research field includes geodetic data processing, navigation, and geodetic coordinate system, etc.

Detailed Contents

List of Abbreviations	XXVII
------------------------------------	-------

Part A Principles of GNSS

1 Introduction to GNSS

<i>Richard B. Langley, Peter J.G. Teunissen, Oliver Montenbruck</i>	3
1.1 Early Satellite Navigation	3
1.2 Concept of GNSS Positioning	5
1.2.1 Ranging Measurements	5
1.2.2 Range-Based Positioning	6
1.2.3 Pseudorange Positioning	7
1.2.4 Precision of Position Solutions	8
1.2.5 GNSS Observation Equations	10
1.3 Modeling the Observations	10
1.3.1 Satellite Orbit and Clock Information	10
1.3.2 Atmospheric Propagation Delay	11
1.4 Positioning Modes	13
1.4.1 Precise Point Positioning	13
1.4.2 Code Differential Positioning	14
1.4.3 Differential Carrier Phase	14
1.5 Current and Developing GNSSs	16
1.5.1 Global Navigation Satellite Systems	16
1.5.2 Regional Navigation Satellite Systems	18
1.5.3 Satellite-Based Augmentation Systems	19
1.6 GNSS for Science and Society at Large	19
References	22

2 Time and Reference Systems

<i>Christopher Jekeli, Oliver Montenbruck</i>	25
2.1 Time	25
2.1.1 Dynamic Time	26
2.1.2 Atomic Time Scales	27
2.1.3 Sidereal and Universal Time, Earth Rotation	27
2.1.4 GNSS System Times	30
2.2 Spatial Reference Systems	31
2.2.1 Coordinate Systems	31
2.2.2 Reference Systems and Frames	34
2.3 Terrestrial Reference System	34
2.3.1 Traditional Geodetic Datums	34
2.3.2 Global Reference System	36
2.3.3 Terrestrial Reference Systems for GNSS Users	39
2.3.4 Frame Transformations	40
2.3.5 Earth Tides	42
2.4 Celestial Reference System	44
2.5 Transformations Between ICRF and ITRF	46

2.5.1	Orientation of the Earth in Space	46
2.5.2	New Conventions	50
2.5.3	Polar Motion	52
2.5.4	Transformations	54
2.6	Perspectives	55
	References	56
3	Satellite Orbits and Attitude	
	<i>Urs Hugentobler, Oliver Montenbruck</i>	59
3.1	Keplerian Motion	59
3.1.1	Basic Properties	59
3.1.2	Keplerian Orbit Model	61
3.1.3	Ground Track and Visibility	63
3.2	Orbit Perturbations	66
3.2.1	Orbit Representation	66
3.2.2	Perturbing Accelerations	67
3.2.3	Perturbations at GNSS Satellite Altitude	71
3.2.4	Radiation Pressure	72
3.2.5	Long-Term Evolution	74
3.2.6	Orbit Accuracy	77
3.3	Broadcast Orbit Models	79
3.3.1	Almanac Models	80
3.3.2	Keplerian Ephemeris Models	81
3.3.3	Cartesian Ephemeris Model	83
3.3.4	Broadcast Ephemeris Generation and Performance	83
3.4	Attitude	85
	References	87
4	Signals and Modulation	
	<i>Michael Meurer, Felix Antriech</i>	91
4.1	Radiofrequency Signals	91
4.1.1	Maxwell's Theory of Electromagnetic Waves and Electromagnetic Foundation	91
4.1.2	Modulation and Complex Baseband Representation of Signals	93
4.1.3	Frequency Bands and Polarization	96
4.2	Spread Spectrum Technique and Pseudo Random Codes	97
4.2.1	Spread Spectrum Signals for Ranging	97
4.2.2	Pseudo-Random Binary Sequences	99
4.2.3	Correlation and Time-Delay Estimation	102
4.3	Modulation Schemes	107
4.3.1	Binary Phase Shift Keying	107
4.3.2	Binary Offset Carrier Modulation and Derivatives	109
4.4	Signal Multiplexing	113
4.4.1	Interplex	114
4.4.2	AltBOC	116
4.5	Navigation Data and Data-Free Channels	117
	References	118

5	Clocks	
	<i>Ron Beard, Ken Senior</i>	121
5.1	Frequency and Time Stability	122
5.1.1	Concepts	123
5.1.2	Characterization of Clock Stability	123
5.2	Clock Technologies	127
5.2.1	Quartz Crystal Oscillators	127
5.2.2	Conventional Atomic Standards	128
5.2.3	Timescale Atomic Standards	135
5.2.4	Small Atomic Clock Technology	136
5.2.5	Developing Clock Technologies	137
5.3	Space-Qualified Atomic Standards	138
5.3.1	Space Rubidium Atomic Clocks	139
5.3.2	Space-Qualified Cesium Beam Clocks	140
5.3.3	Space-Qualified Hydrogen Maser Clocks	141
5.3.4	Space Linear Ion Trap System (LITS)	142
5.3.5	Satellite Onboard Timing Subsystems	142
5.3.6	On-Orbit Performance of Space Atomic Clocks	144
5.4	Relativistic Effects on Clocks	148
5.4.1	Relativistic Terms	148
5.4.2	Coordinate Timescales	150
5.4.3	Geocentric Coordinate Systems	150
5.4.4	Propagation of Signals	153
5.4.5	Relativistic Offset for GNSS Satellite Clocks	154
5.5	International Timescales	155
5.5.1	International Atomic Time (TAI)	155
5.5.2	Coordinated Universal Time (UTC)	157
5.6	GNSS Timescales	158
	References	160
6	Atmospheric Signal Propagation	
	<i>Thomas Hobiger, Norbert Jakowski</i>	165
6.1	Electromagnetic Wave Propagation	165
6.1.1	Maxwell Equations	166
6.1.2	Electromagnetic Wave Propagation in the Troposphere	166
6.1.3	Electromagnetic Wave Propagation in the Ionosphere	167
6.2	Troposphere	168
6.2.1	Characteristics of the Troposphere	168
6.2.2	Tropospheric Refraction	170
6.2.3	Empirical Models of the Troposphere	172
6.2.4	Troposphere Delay Estimation	174
6.3	Ionospheric Effects on GNSS Signal Propagation	177
6.3.1	The Ionosphere	177
6.3.2	Refraction of Transionospheric Radio Waves	179
6.3.3	Diffraction and Scattering of GNSS Signals	183
6.3.4	Ionospheric Models	184
6.3.5	Measurement-Based Ionosphere Correction	189
	References	190

Part B Satellite Navigation Systems

7 The Global Positioning System (GPS) 197

Christopher J. Hegarty 197

7.1 Space Segment 197

7.1.1 Constellation Design and Management..... 197

7.1.2 GPS Satellites..... 199

7.2 Control Segment 203

7.2.1 Overview 203

7.2.2 Evolution of Capabilities 204

7.2.3 Operations..... 204

7.3 Navigation Signals..... 205

7.3.1 Legacy 205

7.3.2 Modernized Signals..... 206

7.3.3 Power Levels 209

7.4 Navigation Data and Algorithms 210

7.4.1 Legacy Navigation (LNAV) Data Overview..... 210

7.4.2 LNAV Error Detection Encoding 211

7.4.3 LNAV Data Content and Related Algorithms 211

7.4.4 Civil Navigation (CNAV) and Civil Navigation-2 (CNAV-2) Data 215

7.5 Time System and Geodesy 216

7.6 Services and Performance 216

References 217

8 GLONASS 219

Sergey Revnivkyh, Alexey Bolkunov, Alexander Serdyukov 219

Oliver Montenbruck..... 219

8.1 Overview 219

8.1.1 History and Evolution 219

8.1.2 Constellation 220

8.1.3 GLONASS Geodesy Reference PZ-90..... 221

8.1.4 GLONASS Time 223

8.2 Navigation Signals and Services..... 225

8.2.1 GLONASS Services..... 225

8.2.2 FDMA Signals 226

8.2.3 CDMA Signals 229

8.3 Satellites 232

8.3.1 GLONASS I/II 233

8.3.2 GLONASS-M 235

8.3.3 GLONASS-K..... 236

8.4 Launch Vehicles 237

8.5 Ground Segment 238

8.6 GLONASS Open Service Performance 241

References 243

9 Galileo 247

Marco Falcone, Jörg Hahn, Thomas Burger 247

9.1 Constellation..... 248

9.2 Signals and Services..... 250

9.2.1 Signal Components and Modulations 251

9.2.2	Navigation Message and Services	256
9.2.3	Ranging Performance	258
9.2.4	Timing Accuracy	263
9.3	Spacecraft	265
9.3.1	Satellite Platform	266
9.3.2	Satellite Payload Description	266
9.3.3	Launch Vehicles	268
9.4	Ground Segment	269
9.5	Summary	270
	References	271
10	Chinese Navigation Satellite Systems	
	<i>Yuanxi Yang, Jing Tang, Oliver Montenbruck</i>	273
10.1	BeiDou Navigation Satellite Demonstration System (BDS-1)	275
10.1.1	System Architecture and Basic Characteristics	275
10.1.2	Navigation Principle	277
10.1.3	Orbit Determination	278
10.1.4	Timing	278
10.2	BeiDou (Regional) Navigation Satellite System (BDS-2)	279
10.2.1	Constellation	279
10.2.2	Signals and Services	281
10.2.3	Navigation Message	283
10.2.4	Space Segment	286
10.2.5	Operational Control System	288
10.2.6	BeiDou Satellite-Based Augmentation System	289
10.2.7	Coordinate Reference System	290
10.2.8	Time System	291
10.3	Performance of BDS-2	293
10.3.1	Service Region	293
10.3.2	Performance of Satellite Clocks	293
10.3.3	Positioning Performance	295
10.3.4	Application Examples	297
10.4	BeiDou (Global) Navigation Satellite System	297
10.5	Brief Introduction of CAPS	298
10.5.1	CAPS Concept and System Architecture	298
10.5.2	Positioning Principle of CAPS	300
10.5.3	Trial CAPS System	301
	References	301
11	Regional Systems	
	<i>Satoshi Kogure, A.S. Ganeshan, Oliver Montenbruck</i>	305
11.1	Concept of Regional Navigation Satellite Systems	306
11.2	Quasi-Zenith Satellite System	306
11.2.1	Overview	306
11.2.2	Constellation	307
11.2.3	Signals and Services	308
11.2.4	Spacecraft	313
11.2.5	Control Segment	317
11.2.6	Operations Concept	319
11.2.7	Current Performance	320

11.3	Indian Regional Navigation Satellite System (IRNSS/NavIC)	321
11.3.1	Constellation	322
11.3.2	Signal and Data Structure	322
11.3.3	Spacecraft	327
11.3.4	Ground Segment	330
11.3.5	System Performance	333
	References	334

12 Satellite Based Augmentation Systems

	<i>Todd Walter</i>	339
12.1	Aircraft Guidance	340
12.1.1	Aviation Requirements	340
12.1.2	Traditional Navigational Aids	341
12.1.3	Receiver Autonomous Integrity Monitoring (RAIM)	341
12.1.4	Satellite-Based Augmentation Systems (SBAS)	342
12.2	GPS Error Sources	343
12.2.1	Satellite Clock and Ephemeris	344
12.2.2	Ionosphere	344
12.2.3	Troposphere	344
12.2.4	Multipath	345
12.2.5	Other Error Sources	345
12.3	SBAS Architecture	345
12.3.1	Reference Stations	345
12.3.2	Master Stations	346
12.3.3	Ground Uplink Stations and Geostationary Satellites	347
12.3.4	Operational Control Centers	349
12.4	SBAS Integrity	349
12.4.1	Integrity Certification	349
12.4.2	Threat Models	350
12.4.3	Overbounding	350
12.5	SBAS User Algorithms	351
12.5.1	Message Structure	351
12.5.2	Message Application	352
12.5.3	Protection Levels	353
12.6	Operational and Planned SBAS Systems	353
12.6.1	Wide Area Augmentation System (WAAS)	353
12.6.2	Multifunction Satellite Augmentation System (MSAS)	356
12.6.3	European Geostationary Navigation Overlay Service (EGNOS)	356
12.6.4	GPS Aided GEO Augmented Navigation (GAGAN)	356
12.6.5	System of Differential Corrections and Monitoring (SDCM)	358
12.6.6	BeiDou Satellite-Based Augmentation System (BDSBAS) ..	358
12.6.7	Korean Augmentation Satellite System (KASS)	358
12.7	Evolution of SBAS	358
12.7.1	Multiple Frequencies	358
12.7.2	Multiple Constellations	359
	References	360

Part C GNSS Receivers and Antennas

13 Receiver Architecture

<i>Bernd Eissfeller, Jong-Hoon Won</i>	365
13.1 Background and History	366
13.1.1 Analog Versus Digital Receivers	366
13.1.2 Early Military Developments	367
13.1.3 Early Civil Developments	368
13.1.4 Early Receiver Developments for Other Satellite Navigation Systems	369
13.1.5 Early BeiDou Receiver Developments	371
13.2 Receiver Building Blocks	372
13.2.1 Antenna	373
13.2.2 RF Front End	376
13.2.3 Analog-to-Digital Conversion	380
13.2.4 Oscillators	383
13.2.5 Chip Technologies	386
13.2.6 Implementation Issues	390
13.3 Multifrequency and Multisystem Receivers	391
13.3.1 Civil Receivers for GPS Modernization	391
13.3.2 Galileo Receivers	394
13.3.3 GLONASS Receivers	394
13.3.4 BeiDou/Compass Receivers	395
13.3.5 Military GPS Receivers	395
13.4 Technology Trends	396
13.4.1 Civil Low-End Trends	396
13.4.2 Civil High-End Trends	396
13.4.3 Trends in Military and/or Governmental Receivers	397
13.5 Receiver Types	397
13.5.1 Navigation Receivers Handheld	397
13.5.2 Navigation Receivers Non-Handheld	397
13.5.3 Engines, OEM Modules, Chips, and Dies	398
13.5.4 Time Transfer Receivers	398
13.5.5 Geodetic Receivers	398
13.5.6 Space Receivers	398
13.5.7 Attitude Determination Receivers	398
References	399

14 Signal Processing

<i>Jong-Hoon Won, Thomas Pany</i>	401
14.1 Overview and Scope	402
14.2 Received Signal Model	403
14.2.1 Generic GNSS Signal	403
14.2.2 Signal Model at RF and IF	404
14.2.3 Correlator Model	404
14.3 Signal Search and Acquisition	406
14.3.1 Test Statistics	406
14.3.2 Acquisition Module Architecture	408
14.3.3 Coherent Integration Methods	409
14.3.4 Search Space	410

14.3.5	Acquisition Performance	411
14.3.6	Handling Data Bits and Secondary Codes	412
14.4	Signal Tracking	413
14.4.1	Architecture	413
14.4.2	Tracking Loop Model	414
14.4.3	Correlators	415
14.4.4	Discriminators	415
14.4.5	Loop Filters	417
14.4.6	NCO and Code/Carrier Generator	419
14.4.7	Aiding	421
14.4.8	Switching Rule	421
14.4.9	BOC Tracking	422
14.4.10	Tracking Performance	422
14.5	Time Synchronization and Data Demodulation	424
14.5.1	Bit/Symbol Synchronization	425
14.5.2	Data Bit/Symbol Demodulation	426
14.5.3	Frame Synchronization	427
14.5.4	Bit Error Correction	427
14.5.5	Data Extraction	428
14.6	GNSS Measurements	428
14.6.1	Code Pseudorange	428
14.6.2	Carrier Phase	431
14.6.3	Doppler	433
14.6.4	Signal Power	434
14.7	Advanced Topics	434
14.7.1	Tracking of GPS P(Y)	434
14.7.2	Generic Data/Pilot Multiplexing Approach	435
14.7.3	Combined Processing of Data and Pilot Signals	436
14.7.4	Combined Processing of Code and Carrier	436
14.7.5	Carrier Tracking Kalman Filter	437
14.7.6	Vector Tracking	438
	References	440
15	Multipath	
	<i>Michael S. Braasch</i>	443
15.1	The Impact of Multipath	444
15.2	Characterizing the Multipath Environment	444
15.3	Multipath Signal Models	448
15.4	Pseudorange and Carrier-Phase Error	450
15.5	Multipath Error Envelopes	450
15.6	Temporal Error Variation, Bias Characteristics and Fast Fading Considerations	453
15.7	Multipath Mitigation	455
15.7.1	Multipath Mitigation via Antenna Placement	455
15.7.2	Antenna Type	456
15.7.3	Receiver Type	457
15.7.4	Measurement Processing	458
15.8	Multipath Measurement	459
15.8.1	Isolation of Pseudorange Multipath	460
15.8.2	Short-Delay Multipath	461

15.8.3	Multipath Repeatability	462
15.8.4	Measurement of Carrier-Phase Multipath	463
15.9	A Note About Multipath Impact on Doppler Measurements	466
15.10	Conclusions	466
	References	466
16	Interference	
	<i>Todd Humphreys</i>	469
16.1	Analysis Technique for Statistically Independent Interference	471
16.1.1	Received Signal Model	471
16.1.2	Thermal-Noise-Equivalent Approximation	471
16.1.3	Limits of Applicability	473
16.1.4	Overview of Interference Effects on Carrier Phase Tracking	474
16.2	Canonical Interference Models	476
16.2.1	Wideband Interference	476
16.2.2	Narrowband Interference	476
16.2.3	Matched-Spectrum Interference	478
16.3	Quantization Effects	479
16.3.1	One-Bit Quantization	479
16.3.2	Multibit Quantization	479
16.4	Specific Interference Waveforms and Sources	481
16.4.1	Solar Radio Bursts	481
16.4.2	Scintillation	482
16.4.3	Unintentional Interference	484
16.4.4	Intentional Interference	485
16.5	Spoofing	485
16.5.1	Generalized Model for Security-Enhanced GNSS Signals ..	486
16.5.2	Attacks Against Security-Enhanced GNSS Signals	486
16.6	Interference Detection	491
16.6.1	C/N_0 Monitoring	491
16.6.2	Received Power Monitoring	491
16.6.3	Augmented Received Power Monitoring	493
16.6.4	Spectral Analysis	494
16.6.5	Cryptographic Spoofing Detection	495
16.6.6	Antenna-Based Techniques	497
16.6.7	Innovations-Based Techniques	497
16.7	Interference Mitigation	498
16.7.1	Spectrally or Temporally Sparse Interference	498
16.7.2	Spectrally and Temporally Dense Interference	499
16.7.3	Antenna-Based Techniques	500
	References	501
17	Antennas	
	<i>Moazam Maqsood, Steven Gao, Oliver Montenbruck</i>	505
17.1	GNSS Antenna Characteristics	506
17.1.1	Center Frequency	507
17.1.2	Bandwidth	507
17.1.3	Radiation Pattern	507
17.1.4	Antenna Gain	507
17.1.5	3 dB Beam Width	507

17.1.6	Polarization	508
17.1.7	Axial Ratio	508
17.1.8	Impedance Matching and Return Loss	508
17.1.9	Front-to-Back and Multipath Ratio	509
17.1.10	Phase-Center Stability	509
17.2	Basic GNSS Antenna Types	509
17.2.1	Microstrip Patch Antenna	509
17.2.2	Helix Antenna	510
17.2.3	Quadrifilar Helix Antenna	511
17.2.4	Spiral Antenna	512
17.2.5	Wide-Band Bow-Tie Turnstile Antenna	513
17.2.6	Wide-Band Pinwheel Antenna	513
17.3	Application-Specific GNSS Antennas	513
17.3.1	Hand-Held Terminals	513
17.3.2	Surveying and Geodesy	514
17.3.3	Aviation	515
17.3.4	Space Applications	516
17.3.5	Antijamming Antennas	517
17.3.6	GNSS Remote Sensing	518
17.4	Multipath Mitigation	519
17.4.1	Metallic Reflector Ground Plane	520
17.4.2	Choke-Ring Ground Plane	520
17.4.3	Noncutoff Corrugated Ground Plane	521
17.4.4	Convex Impedance Ground Plane	521
17.4.5	3-D Choke-Ring Ground Plane	521
17.4.6	Cross Plate Reflector Ground Plane	522
17.4.7	Electromagnetic Band Gap (EBG) Substrate	522
17.5	Antennas for GNSS Satellites	523
17.5.1	Concentric Helix Antenna Arrays	523
17.5.2	Patch Antenna Arrays	524
17.5.3	Reflector-Backed Monofilar Antenna	526
17.6	Antenna Measurement and Calibration	527
17.6.1	Basic Antenna Testing	527
17.6.2	Phase-Center Calibration	528
	References	531

18 Simulators and Test Equipment

	<i>Mark G. Petovello, James T. Curran</i>	535
18.1	Background	537
18.1.1	Received RF Signal	537
18.1.2	GNSS Receivers	540
18.1.3	GNSS Simulators	540
18.1.4	Record and Playback Systems	542
18.1.5	Details	543
18.2	RF-Level Simulators	543
18.2.1	Implementation	544
18.2.2	Important Considerations	544
18.3	IF-Level Simulators	546
18.3.1	Implementation	547
18.3.2	Important Considerations	548

18.4	Record and Playback Systems	549
18.4.1	Implementation	550
18.4.2	Important Considerations.....	550
18.5	Measurement-Level Simulators	552
18.5.1	Implementation	553
18.5.2	Important Considerations.....	553
18.6	Combining Live and Simulated Data	554
18.6.1	Implementation	555
18.6.2	Important Considerations.....	555
18.7	Other Considerations	556
18.7.1	GNSS Systems Supported	556
18.7.2	Interference and Spoofing	556
18.7.3	Other Data	556
18.7.4	Configurability	556
18.7.5	Expandability	556
18.8	Summary	557
	References	557

Part D GNSS Algorithms and Models

19	Basic Observation Equations	
	<i>André Hauschild</i>	561
19.1	Observation Equations	561
19.1.1	Pseudorange Measurements	561
19.1.2	Carrier-Phase Measurements	563
19.1.3	Doppler Measurements	563
19.2	Relativistic Effects	564
19.3	Atmospheric Signal Delays	565
19.3.1	Ionosphere	566
19.3.2	Troposphere	568
19.4	Carrier-Phase Wind-Up	569
19.4.1	Wind-Up Effect for Radio Waves	569
19.4.2	GNSS Satellite Attitude Modeling	570
19.5	Antenna Phase-Center Offset and Variations	572
19.5.1	Overview	573
19.5.2	Calibration Techniques	574
19.5.3	Examples for Phase-Center Variations	575
19.6	Signal Biases	576
19.6.1	Pseudorange Biases	576
19.6.2	Carrier-Phase Biases	578
19.7	Receiver Noise and Multipath	578
19.7.1	Receiver Noise	578
19.7.2	Multipath Errors	579
	References	579
20	Combinations of Observations	
	<i>André Hauschild</i>	583
20.1	Fundamental Equations	583
20.2	Combinations of Single-Satellite and Single-Receiver Observations	586
20.2.1	Narrow- and Wide-Lane Combinations	586

20.2.2	Ionosphere Combination	589
20.2.3	Ionosphere-Free Combination	590
20.2.4	Multipath Combination	591
20.3	Combinations of Multisatellite and Multireceiver Observations	594
20.3.1	Between-Receiver Single Difference	594
20.3.2	Between-Satellite Single Difference	596
20.3.3	Double Difference	596
20.3.4	Triple Difference	598
20.3.5	Single and Double Difference on Zero-Baselines	599
20.4	Pseudorange Filtering	601
	References	603

21 Positioning Model

	<i>Dennis Odijk</i>	605
21.1	Nonlinear Observation Equations	606
21.1.1	Single-GNSS Observation Equations	606
21.1.2	Multi-GNSS Observation Equations	607
21.2	Linearization of the Observation Equations	609
21.2.1	Linearizing the Receiver-Satellite Range	609
21.2.2	Linearized Observation Equations	612
21.3	Point Positioning Models	612
21.3.1	Computation of the Satellite Clocks and Hardware Code (Group) Delays	613
21.3.2	Some Remarks on the TGDs/DCBs	615
21.3.3	Computation/Estimation of the Atmospheric Errors	615
21.3.4	Single-Constellation SPP Model	615
21.3.5	Multiconstellation SPP Model	617
21.3.6	Precision and DOP	618
21.3.7	PPP Model	619
21.4	Relative Positioning Models	623
21.4.1	Principle of DGNSS and (PPP-)RTK	623
21.4.2	Impact of Orbit Errors	625
21.4.3	Ionosphere-Fixed/Weighted/Float Models	625
21.4.4	Undifferenced Relative Positioning Models	625
21.4.5	PPP-RTK Models	627
21.4.6	Link Between PPP-RTK and PPP	630
21.5	Differenced Positioning Models	631
21.5.1	Single Differencing	631
21.5.2	Double and Triple Differencing	632
21.5.3	Redundancy of the Differenced Models	633
21.6	The Positioning Concepts Related	633
21.6.1	Global Positioning: SPP/PPP	633
21.6.2	Regional Positioning: Network DGNSS/RTK	634
21.6.3	Local Positioning: Single-Baseline DGNSS/RTK	634
21.6.4	Global/Regional Positioning: PPP-RTK	634
21.6.5	Accuracy of the Positioning Concepts	635
	References	635

22 Least-Squares Estimation and Kalman Filtering

<i>Sandra Verhagen, Peter J.G. Teunissen</i>	639
22.1 Linear Least-Squares Estimation	639
22.1.1 Least-Squares Principle	639
22.1.2 Weighted Least-Squares	640
22.1.3 Computation of LS Solution	640
22.1.4 Statistical Properties	641
22.2 Optimal Estimation	641
22.2.1 Best Linear Unbiased Estimation	641
22.2.2 Maximum Likelihood Estimation	642
22.2.3 Confidence Regions	642
22.3 Special Forms of Least Squares	644
22.3.1 Recursive Estimation	644
22.3.2 Estimation with Partitioned Parameter Vector	646
22.3.3 Block Estimation	647
22.3.4 Constrained Least-Squares	647
22.3.5 Rank-Defect Least Squares	647
22.3.6 Non-Linear Least-Squares	648
22.4 Prediction and Filtering	650
22.4.1 Prediction Problem	650
22.4.2 Minimum Mean Squared Error Prediction	651
22.4.3 Properties of MMSE Prediction	653
22.5 Kalman Filtering	653
22.5.1 Model Assumptions	653
22.5.2 The Kalman Filter Recursion	654
22.5.3 Kalman Filter Information Form	655
22.5.4 Extended Kalman Filter	656
22.5.5 Smoothing	657
References	659

23 Carrier Phase Integer Ambiguity Resolution

<i>Peter J.G. Teunissen</i>	661
23.1 GNSS Ambiguity Resolution	662
23.1.1 The GNSS Model	662
23.1.2 Ambiguity Resolution Steps	662
23.1.3 Ambiguity Resolution Quality	663
23.2 Rounding and Bootstrapping	666
23.2.1 Integer Rounding	666
23.2.2 Vectorial Rounding	666
23.2.3 Integer Bootstrapping	667
23.2.4 Bootstrapped Success Rate	668
23.3 Linear Combinations	669
23.3.1 Z-transformations	669
23.3.2 (Extra) Widelaning	669
23.3.3 Decorrelating Transformation	670
23.3.4 Numerical Example	672
23.4 Integer Least-Squares	673
23.4.1 Mixed Integer Least-Squares	673
23.4.2 The ILS Computation	674
23.4.3 Least-Squares Success Rate	676

23.5	Partial Ambiguity Resolution	677
23.6	When to Accept the Integer Solution?	678
23.6.1	Model- and Data-Driven Rules	678
23.6.2	Four Ambiguity Resolution Steps	679
23.6.3	Quality of Accepted Integer Solution	679
23.6.4	Fixed Failure-Rate Ratio Test	680
23.6.5	Optimal Integer Ambiguity Test	681
	References	683

24 Batch and Recursive Model Validation

	<i>Peter J.G. Teunissen</i>	687
24.1	Modeling and Validation	687
24.2	Batch Model Validation	689
24.2.1	Null versus Alternative Hypothesis	689
24.2.2	Unbiased versus Biased Solution	689
24.2.3	Effect of the Influential Bias	690
24.3	Testing for a Bias	692
24.3.1	The Most Powerful Test Statistic	692
24.3.2	Alternative Expressions for Test Statistic T_q	695
24.3.3	Test Statistic T_q Expressed in LS Residuals	696
24.3.4	Optimality of the w -Test Statistic	697
24.3.5	The Minimal Detectable Bias	697
24.3.6	Hazardous Missed Detection	700
24.4	Testing Procedure	705
24.4.1	Detection, Identification and Adaptation	705
24.4.2	Data Snooping	706
24.4.3	Unknowns in the Stochastic Model	709
24.5	Recursive Model Validation	710
24.5.1	Model and Filter	710
24.5.2	Models and UMPI Test Statistic	711
24.5.3	Local and Global Testing	711
24.5.4	Recursive Detection	712
24.5.5	Recursive Identification	713
24.5.6	Recursive Adaptation: General Case	714
24.5.7	Recursive Adaptation: Special GNSS Case	715
	References	717

Part E Positioning and Navigation

25 Precise Point Positioning

	<i>Jan Kouba, François Lahaye, Pierre Tétreault</i>	723
25.1	PPP Concept	724
25.1.1	Observation Equations	724
25.1.2	Adjustment and Quality Control	725
25.2	Precise Positioning Correction Models	726
25.2.1	Atmospheric Propagation Delays	728
25.2.2	Antenna Effects	730
25.2.3	Site Displacement Effects	732
25.2.4	Differential Code Biases	733
25.2.5	Compatibility and Conventions	734

25.3	Specific Processing Aspects	735
25.3.1	Single-Frequency Positioning	735
25.3.2	GLONASS PPP Considerations	736
25.3.3	New Signals and Constellations	737
25.3.4	Phase Ambiguity Fixing in PPP	739
25.4	Implementations	741
25.4.1	Post-Processed Solutions	741
25.4.2	Real-Time Solutions	742
25.4.3	PPP Positioning Services	742
25.5	Examples	743
25.5.1	Static PPP Solutions	743
25.5.2	Kinematic PPP Solutions	743
25.5.3	Tropospheric Zenith Path Delay	745
25.5.4	Station Clock Solutions	745
25.6	Discussion	746
	References	747
26	Differential Positioning	
	<i>Dennis Odijk, Lambert Wanninger</i>	753
26.1	Differential GNSS: Concepts	753
26.1.1	Differential GNSS Observation Equations	753
26.1.2	Differential GNSS Biases	754
26.2	Differential Navigation Services	760
26.2.1	DGNSS Implementations	760
26.2.2	DGNSS Services	761
26.2.3	Data Communication: RTCM Message	762
26.2.4	Latency of DGNSS Corrections	762
26.3	Real-Time Kinematic Positioning	763
26.3.1	Double-Differenced Positioning Model	763
26.3.2	Carrier-Phase-Based Positioning Methods	764
26.3.3	GLONASS RTK Positioning	766
26.3.4	Multi-GNSS RTK Positioning	768
26.3.5	RTK Positioning Examples	770
26.4	Network RTK	774
26.4.1	From RTK to Network RTK	774
26.4.2	Data Processing Methods for Network RTK	774
26.4.3	Network RTK Correction Models	776
26.4.4	Refined Virtual Reference Stations	777
26.4.5	From Network RTK to PPP-RTK	778
	References	778
27	Attitude Determination	
	<i>Gabriele Giorgi</i>	781
27.1	Six Degrees of Freedom	781
27.2	Attitude Parameterization	784
27.2.1	The Space of Rotations	784
27.2.2	Parameterization of the Rotation Matrix	784
27.3	Attitude Estimation from Baseline Observations	787
27.3.1	Estimation of the Orthonormal Matrix of Rotations	787
27.3.2	Orthogonal Procrustes Problem	788

27.3.3	Weighted Orthogonal Procrustes Problem	789
27.3.4	Attitude Estimation with Fully Populated Weight Matrix..	789
27.3.5	On the Precision of Attitude Estimation	790
27.4	The GNSS Attitude Model.....	790
27.4.1	Potential Model Errors and Misspecification	791
27.4.2	Resolution of the GNSS Attitude Model	792
27.4.3	The GNSS Ambiguity and Attitude Estimation	793
27.4.4	The Quality of Ambiguity and Attitude Estimations.....	795
27.5	Applications.....	798
27.5.1	Space Operations	798
27.5.2	Aeronautics Applications.....	800
27.5.3	Marine Navigation	802
27.5.4	Land Applications	803
27.6	An Overview of GNSS/INS Sensor Fusion for Attitude Determination	804
	References	806
28	GNSS/INS Integration	
	<i>Jay A. Farrell, Jan Wendel</i>	811
28.1	State Estimation Objectives	812
28.2	Inertial Navigation	813
28.2.1	Problem Statement	813
28.2.2	Sensor Models	813
28.2.3	INS Computations	813
28.2.4	INS Error State	814
28.2.5	Performance Characterization	814
28.3	Inertial Sensors	815
28.3.1	Gyroscopes	815
28.3.2	Accelerometers.....	816
28.3.3	Inertial Sensor Errors	816
28.4	Strapdown Inertial Navigation	818
28.4.1	Coordinate Systems.....	818
28.4.2	Attitude Calculations	819
28.4.3	Velocity Calculations.....	821
28.4.4	Position Calculations	821
28.5	Analysis of Error Effects	822
28.5.1	Short-Term Effects	822
28.5.2	Long-Term Effects	823
28.6	Aided Navigation	824
28.7	State Estimation	824
28.8	GNSS and Aided INS	825
28.8.1	Loose (Position Domain) Coupling	825
28.8.2	Tight (Observable Domain) Coupling	826
28.8.3	Ultra-Tight or Deep Coupling	826
28.8.4	Illustrative Comparison.....	828
28.9	Detailed Example.....	828
28.9.1	System Model	828
28.9.2	Measurement Models	831
28.10	Alternative Estimation Methods.....	835
28.10.1	Standalone GNSS.....	835
28.10.2	Advanced Bayesian Estimation	837

28.11 Looking Forward	838
References	839
29 Land and Maritime Applications	
<i>Allison Kealy, Terry Moore</i>	841
29.1 Land-Based Applications of GNSS	842
29.1.1 Personal Devices	843
29.1.2 Location-Based Services	845
29.1.3 Positioning Technologies and Techniques for PN and LBS	846
29.1.4 Intelligent Transport Systems	853
29.2 Rail Applications	856
29.2.1 Signaling and Train Control	857
29.2.2 Freight and Fleet Management	862
29.2.3 Passenger Information Systems	863
29.3 Maritime Applications	863
29.3.1 GNSS Performance Requirements for Maritime Applications	864
29.3.2 Maritime Navigation	867
29.3.3 eLoran	869
29.3.4 Automatic Identification System	869
29.3.5 Shipping Container Tracking	872
29.4 Outlook	873
References	873
30 Aviation Applications	
<i>Richard Farnworth</i>	877
30.1 Overview	878
30.1.1 Conventional Navigation	878
30.1.2 Area Navigation – RNAV	879
30.1.3 The Arrival of GNSS	880
30.2 Standardising GNSS for Aviation	881
30.2.1 Aircraft Based Augmentation Systems	882
30.2.2 Satellite Based Augmentation Systems	882
30.2.3 Ground Based Augmentation Systems	883
30.3 Evolution of the Flight Deck	884
30.3.1 The Navigation Data Chain	885
30.3.2 General Aviation	885
30.3.3 Helicopters	885
30.4 From the RNP Concept to PBN	886
30.4.1 Performance Based Navigation (PBN)	886
30.4.2 Navigation Specifications	886
30.5 GNSS Performance Requirements	888
30.5.1 Description of the Relevant Parameters	888
30.5.2 GNSS Integrity Concepts	890
30.6 Linking the PBN Requirements and the GNSS Requirements	891
30.6.1 Phases of Flight	891
30.6.2 RNAV Approaches	893
30.6.3 RNP AR APCH	895
30.7 Flight Planning and NOTAMS	897

30.8	Regulation and Certification	897
30.8.1	Airworthiness Certification.....	897
30.8.2	Operational Approvals.....	898
30.9	Military Aviation Applications	898
30.10	Other Aviation Applications of GNSS	899
30.10.1	Surveillance (ADS-B).....	899
30.10.2	Datalink	899
30.11	Future Evolution	900
30.11.1	GNSS Vulnerability and Alternative-PNT	900
30.11.2	Rationalisation of the Navigation Infrastructure.....	900
30.11.3	Multi-Constellation.....	901
	References	901
31	Ground Based Augmentation Systems	905
	<i>Sam Pullen</i>	905
31.1	Components	906
31.2	An Overview of Local Area Approaches	907
31.2.1	Pseudorange Corrections	907
31.2.2	Carrier-Phase Corrections	908
31.2.3	Reference Station Distribution.....	908
31.2.4	Broadcast Techniques	908
31.3	Ground-Based Augmentation Systems	909
31.3.1	Overview and Requirements.....	909
31.3.2	Generation of Differential Corrections	910
31.3.3	Fault Monitoring.....	911
31.3.4	User Processing and Integrity Verification	917
31.3.5	Additional Threats: RF Interference and Ionosphere	921
31.3.6	Equipment and Siting Considerations	925
31.3.7	Typical GBAS Errors and Protection Levels.....	926
31.3.8	Existing GBAS Ground Systems and Airborne Equipment..	928
31.4	Augmentation via Ranging Signals Pseudolites.....	928
31.4.1	Origins and Use in Local-Area DGNSS	928
31.4.2	New-Generation Pseudolite Systems for Commercial Applications	929
31.5	Outlook.....	930
	References	930
32	Space Applications	933
	<i>Oliver Montenbruck</i>	933
32.1	Flying High	933
32.1.1	GNSS Tracking in Space	934
32.1.2	Spaceborne GPS Receivers	936
32.2	Spacecraft Navigation	938
32.2.1	Trajectory Models	939
32.2.2	Real-Time Navigation	942
32.2.3	Precise Orbit Determination	946
32.3	Formation Flying and Rendezvous.....	951
32.3.1	Differential Observations and Models.....	952
32.3.2	Estimation Concepts.....	954
32.3.3	Ambiguity Resolution	955
32.3.4	Flight Demonstrations.....	955

32.4	Other Applications.....	957
32.4.1	Attitude Determination	957
32.4.2	Ballistic Missions	958
32.4.3	GNSS Radio Science	959
	References	959

Part F Surveying, Geodesy and Geodynamics

33	The International GNSS Service	
	<i>Gary Johnston, Anna Riddell, Grant Hausler</i>	967
33.1	Mission and Organization	967
33.1.1	Mission	967
33.1.2	Structure.....	968
33.2	Components	969
33.2.1	IGS Governing Board and Executive Committee.....	969
33.2.2	IGS Central Bureau	969
33.2.3	IGS Network	970
33.2.4	Analysis Centers.....	970
33.2.5	Data Centers (DCs)	970
33.2.6	Working Groups.....	971
33.3	IGS Products	972
33.3.1	Orbits and Clocks	972
33.3.2	Earth Orientation and Site Coordinates	973
33.3.3	Atmospheric Parameters	974
33.3.4	Biases.....	975
33.4	Pilot Projects and Experiments.....	976
33.4.1	Real-Time	976
33.4.2	Multi-GNSS	978
33.5	Outlook.....	981
	References	981
34	Orbit and Clock Product Generation	
	<i>Jan P. Weiss, Peter Steigenberger, Tim Springer</i>	983
34.1	Global Tracking Network	984
34.2	Models	985
34.2.1	Reference Frame Transformation	985
34.2.2	Site Displacement Effects	985
34.2.3	Tropospheric Delay	988
34.2.4	Ionospheric Delay	988
34.2.5	Relativistic Effects	989
34.2.6	Antenna Phase Center Calibrations	989
34.2.7	Phase Wind-Up	989
34.2.8	GNSS Transmitter Models and Information	990
34.2.9	Models in Downstream Applications	991
34.3	POD Process	992
34.4	Estimation Strategies.....	993
34.4.1	Estimators	993
34.4.2	Parameterization	994
34.4.3	Ground Stations	995
34.4.4	GNSS Orbits	995

34.4.5	Clock Offsets	996
34.4.6	Earth Orientation	996
34.4.7	Phase Ambiguity Resolution	996
34.4.8	Multi-GNSS Processing	997
34.4.9	Terrestrial Reference Frame	998
34.4.10	Sample Parameterizations	998
34.4.11	Reducing Computation Cost	999
34.5	Software	1000
34.6	Products	1001
34.6.1	IGS Orbit and Clock Combination	1002
34.6.2	Formats and Transmission	1004
34.6.3	Using Products	1005
34.7	Outlook	1005
	References	1006

35 Surveying

	<i>Chris Rizos</i>	1011
35.1	Precise Positioning Techniques	1013
35.1.1	Static Positioning	1014
35.1.2	Rapid-Static Positioning	1016
35.1.3	Kinematic Positioning	1017
35.1.4	Real-Time Differential GNSS Positioning	1019
35.1.5	Precise Point Positioning	1021
35.2	Geodetic and Land Surveying	1023
35.2.1	Geodetic Survey Applications	1023
35.2.2	Land Surveying Operations	1024
35.2.3	Land Surveying and Mapping Applications	1027
35.3	Engineering Surveying	1029
35.3.1	Engineering Surveying Real-Time Operations	1029
35.3.2	Engineering Surveying Applications	1030
35.3.3	Project Execution and Related Issues	1032
35.4	Hydrographic Surveying	1033
35.4.1	Hydrographic Surveying Applications	1033
35.4.2	Operational Issues	1035
	References	1035

36 Geodesy

	<i>Zuheir Altamimi, Richard Gross</i>	1039
36.1	GNSS and IAG's Global Geodetic Observing System	1039
36.1.1	The International Association of Geodesy	1040
36.1.2	The Global Geodetic Observing System	1041
36.2	Global and Regional Reference Frames	1044
36.2.1	Reference Frame Representations for the Deformable Earth	1044
36.2.2	Global Terrestrial Reference Frames	1047
36.2.3	GNSS-Based Reference Frames and Their Relationship with the ITRF	1050
36.2.4	General Guidelines for GNSS-Based Reference Frame Implementation	1052
36.2.5	GNSS, Reference Frame and Sea Level Monitoring	1053

36.3	Earth Rotation, Polar Motion, and Nutation	1054
36.3.1	Theory of the Earth's Rotation	1055
36.3.2	Length-of-Day	1055
36.3.3	Polar Motion	1056
36.3.4	Nutation	1058
	References	1059
37	Geodynamics	
	<i>Jeff Freymueller</i>	1063
37.1	GNSS for Geodynamics	1064
37.1.1	Accuracy Requirements	1064
37.1.2	Today's GNSS Accuracy	1065
37.1.3	Accuracy Limitations and Error Sources	1066
37.2	History and Establishment of GNSS Networks for Geodynamics	1067
37.2.1	Campaign GPS Networks	1067
37.2.2	Continuous GNSS Networks for Geodynamics	1068
37.2.3	The Importance of Global Networks	1071
37.3	Rigid Plate Motions	1071
37.4	Plate Boundary Deformation and the Earthquake Cycle	1073
37.4.1	Plate Boundary Zones	1074
37.4.2	Earthquake Cycle Deformation	1075
37.4.3	Elastic Block Modeling	1077
37.5	Seismology	1078
37.5.1	Static Displacements	1079
37.5.2	Dynamic Displacements from Kinematic GNSS	1082
37.5.3	Real-Time Application to Earthquake Warning and Tsunami Warning	1083
37.5.4	Transient Slip	1085
37.5.5	Postseismic Deformation	1085
37.6	Volcano Deformation	1088
37.7	Surface Loading Deformation	1091
37.7.1	Computing Loading Displacements	1091
37.7.2	Examples of Loading Displacements in GNSS Studies	1092
37.7.3	Loads and Load Models	1093
37.7.4	Impacts of Loading Variations on Reference Frame	1094
37.7.5	Glacial Isostatic Adjustment (GIA)	1095
37.8	The Multi-GNSS Future	1099
	References	1100

Part G GNSS Remote Sensing and Timing

38	Monitoring of the Neutral Atmosphere	
	<i>Gunnar Elgered, Jens Wickert</i>	1109
38.1	Ground-Based Monitoring of the Neutral Atmosphere	1110
38.1.1	Accuracy of Propagation Delays	1111
38.1.2	From Delays to Water Vapor Content	1112
38.1.3	Applications to Weather Forecasting	1115
38.1.4	Applications to Climate Research	1118

38.2	GNSS Radio Occultation Measurements	1120
38.2.1	Introduction and History	1120
38.2.2	Basic Principles and Data Analysis	1120
38.2.3	Occultation Missions	1124
38.2.4	Occultation Number and Global Distribution	1125
38.2.5	Measurement Accuracy	1126
38.2.6	Prospects of New Navigation Satellite Systems	1127
38.2.7	Weather Prediction	1128
38.2.8	Climate Monitoring	1128
38.2.9	Synergy of GNSS Radio Occultation with Reflectometry	1131
38.3	Outlook	1132
	References	1133
39	Ionosphere Monitoring	
	<i>Norbert Jakowski</i>	1139
39.1	Ground-Based GNSS Monitoring	1140
39.1.1	Calibration of TEC Measurements	1140
39.1.2	Global Ionosphere Maps	1141
39.2	Space-Based GNSS Monitoring	1144
39.2.1	GNSS Radio Occultation	1144
39.2.2	Ionosphere/Plasmasphere Reconstruction	1145
39.3	GNSS-Based 3-D-Tomography	1147
39.3.1	Reconstruction Techniques	1147
39.3.2	Near-Real-Time Reconstruction	1148
39.4	Scintillation Monitoring	1148
39.4.1	Climatology of Radio Scintillations Deduced from GNSS Observations	1148
39.4.2	Scintillation Measurement Networks	1151
39.5	Space Weather	1152
39.5.1	Direct Impact of Solar Radiation and Energetic Particles ..	1152
39.5.2	Ionospheric Perturbations and Associated Effects	1153
39.5.3	Prediction of Space Weather Phenomena	1155
39.6	Coupling with Lower Geospheres	1156
39.6.1	Atmospheric Signatures	1156
39.6.2	Earthquake Signatures	1158
39.7	Information and Data Services	1159
	References	1159
40	Reflectometry	
	<i>Antonio Rius, Estel Cardellach</i>	1163
40.1	Receivers	1164
40.1.1	GNSS-R Receivers	1165
40.2	Models	1167
40.2.1	Delay-Doppler Coordinates	1167
40.2.2	The Ambiguity Function	1167
40.2.3	The Noiseless Waveform Model	1169
40.2.4	Floor Noise Model	1169
40.2.5	Maximum Coherence Averaging Interval	1170
40.2.6	Speckle Noise	1171
40.2.7	Observed versus Modeled Waveforms	1171

40.3	Applications.....	1172
40.3.1	Sea Surface Altimetry.....	1173
40.3.2	Sea Surface Scatterometry.....	1175
40.3.3	Sea Surface Permittivity.....	1177
40.3.4	Cryosphere: Ice and Snow.....	1177
40.3.5	Land: Soil Moisture and Vegetation.....	1181
40.4	Spaceborne Missions.....	1182
	References	1183
41	GNSS Time and Frequency Transfer	
	<i>Pascale Defraigne</i>	1187
41.1	GNSS Time and Frequency Dissemination.....	1187
41.1.1	Getting UTC from GNSS.....	1188
41.1.2	GNSS Disciplined Oscillators.....	1189
41.2	Remote Clock Comparisons.....	1191
41.2.1	The GNSS Time Transfer Technique.....	1191
41.2.2	Time Transfer Standard CGGTTS.....	1192
41.2.3	Common View or All-in-View.....	1193
41.2.4	Precise Point Positioning.....	1194
41.3	Hardware Architecture and Calibration.....	1197
41.3.1	Time Receivers.....	1197
41.3.2	Hardware Calibration.....	1198
41.4	Multi-GNSS Time Transfer.....	1201
41.4.1	General Requirements.....	1201
41.4.2	GPS + GLONASS Combination.....	1202
41.4.3	Time Transfer with Galileo and BeiDou.....	1203
41.5	Conclusions.....	1203
	References	1204
	Annex A: Data Formats	1207
	Annex B: GNSS Parameters	1233
	About the Authors	1241
	Detailed Contents	1251
	Glossary of Defining Terms	1253
	Subject Index	1281

Glossary of Defining Terms

A

Accuracy

A measure for the closeness of a measured or estimated quantity to the quantity's true value.

Acquisition

The process carried out by a GNSS receiver to identify which GNSS signals are present in the received signal and to determine an approximate code delay and Doppler shift of those signals.

Aircraft Based Augmentation System (ABAS)

A GNSS augmentation system that augments and/or integrates the information obtained from the other GNSS elements with information available on board the aircraft.

Airport Pseudolite (APL)

A ► *pseudolite* located within the boundary of an airport designed to augment the positioning geometry of aircraft approaching or traveling near that airport.

Albedo

A measure for the reflectivity of a body. Earth albedo is a source of indirect radiation pressure acting on a satellite (► *Earth Radiation Pressure*).

Alert

A timely warning that a system may no longer be operating as previously described and that it could now be providing misleading information.

Alert Limit

A maximum tolerable positioning error for an operation to safely proceed. An alert limit has an associated ► *integrity* risk probability and a maximum time before the user must be notified when the alert limit cannot be assured to that integrity risk level.

Allan Deviation (ADEV)

The square-root of the ► *Allan Variance*.

Allan Variance (AVAR)

A measure of clock stability named after the physicist David W. Allan. It describes the variance of the average clock frequency over different time scales not taking into account constant frequency errors.

Almanac

A set of coarse orbit parameters for an entire GNSS constellation that is transmitted as part of the ► *navigation message* by each satellite of the constellation. The almanac facilitates rapid acquisition of all visible satellites.

Alternative BOC (AltBOC)

A modulation scheme combining two quadrature phase shift key signals in adjacent frequency bands, into a combined wideband signal with superior ► *noise* and ► *multipath* properties. AltBOC modulation was first employed for the Galileo E5a/E5b signal.

Ambiguity Dilution of Precision (ADOP)

A scalar measure in cycles that captures the intrinsic precision of the estimated float ambiguity vector. The ADOP is invariant for ambiguity re-parameterizing ► *Z-transformations* and facilitates easy-to-compute approximations of the ambiguity success rate.

Ambiguity

The initial unknown offset in a carrier phase observation when a GNSS receiver first locks onto a GNSS signal. It is the sum of the initial phase and the integer ambiguity. The ambiguity value remains constant as long as the receiver remains locked on the signal.

Ambiguity Fixed Solution

An integer ambiguity resolved GNSS parameter solution. The precision of the ambiguity fixed solution is never poorer than that of the ► *ambiguity float solution*.

Ambiguity Float Solution

A GNSS parameter solution for which the ambiguities are not resolved as integers. The precision of the ambiguity float solution is never better than that of the ► *ambiguity fixed solution*.

Ambiguity Resolution

► *Integer ambiguity resolution*

Ambiguity Success Rate

The probability of correct ► *integer ambiguity* estimation.

Anechoic Chamber

A room or cabinet, the interior of which is non-reflective, absorbing radio-frequency signals originating from within the chamber. In many cases it is often desirable that it also contains these signals and isolates the interior from externally originating signals. Typically used when it is necessary to model an infinite, reflector free space, when conducting a broadcast test, or in cases where the test must be isolated from external interference, or when the test involves the broadcast of signals in protected or restricted bands.

Antenna

A hardware unit that converts electrical energy to electromagnetic waves or vice versa.

Antenna Gain Pattern

A direction-dependent measure of the antenna's efficiency to convert electrical power into radio waves (transmit antenna) or vice versa (receiving antenna).

Antenna Phase Center

The point in the antenna radiation pattern where all the field emanates from or converges to.

Antenna Reference Point

A well defined and easily accessible mechanical point of the antenna to which the electrical ► *antenna*

phase center can be referred and which can itself be referred to the marker of a geodetic monument.

Antenna Thrust

The recoil force due to transmission of GNSS microwave signals which causes a non-gravitational acceleration of the satellite.

ANTEX

The antenna exchange format is used by the IGS to distribute a consistent set of absolute ► *antenna phase center* corrections for both GNSS receivers and satellites. Includes both ► *phase center offsets* and ► *phase center variations*.

Anti-spoofing (A/S)

The use of signal encryption for the military ► *P-code* of the ► *Global Positioning System* to avoid the generation and transmission of spoofed signals by an enemy (► *spoofing*).

Apogee

The most distant point of a satellite orbit from the Earth.

Apogee Kick Motor

An onboard rocket that is used to place a satellite in its final orbit from a highly elliptical transfer orbit.

Approach Procedure with Vertical guidance (APV)

An instrument approach procedure with both lateral and vertical guidance supporting operations with a performance that is between non-precision and precision approach.

Area Correction Parameters

► *Flächenkorrekturparameter*

Area Navigation (RNAV)

RNAV is a method of navigation which permits the operation of an aircraft on any desired flight path. It allows its position to be continuously determined wherever it is rather than only along tracks between individual ground-based navigation aids.

Ascending Node

The point on an orbit at which a satellite crosses the Equator from south to north.

Astronomical Unit

A unit of length that is used to specify distances within the Solar System. One astronomical unit (AU) is approximately equal to the mean distance between the Earth and the Sun, and amounts to roughly 149.6 Mio. km.

Atmosphere

The shells of gases that surround the Earth. The atmosphere affects the motion of satellites in low Earth orbit (► *drag*) and the radio propagation of electromagnetic waves such as radio navigation signals (► *troposphere*, ► *ionosphere*).

Atomic Clock

A device using the frequency of an electronic transition of atoms to generate a timescale. Atomic clocks are the most stable time and frequency standards known.

Atomic Fountain

A signal generator/frequency standard based on the laser cooling of atoms in a magneto-optical trap and once cooled to near absolute zero conditions, they are

launched upwards in the gravity field in a *fountain* arrangement. In this manner the atoms pass through an interrogation region that stimulates a hyperfine transition of the atom's ground state.

Atomic Frequency Standard

A signal generator based on the interrogation of the change in a particular energy state of a specific atom contained in a controlled environment.

Atomic Time Scale

A time scale based on atomic or molecular resonance phenomena. Elapsed time is measured by counting cycles of a frequency locked to an atomic or molecular transition. Atomic time scales differ from the earlier astronomical time scales, which define the second based on the rotation of the Earth on its axis. ► *Coordinated Universal Time* (UTC) is an atomic time scale, since it defines the second based on the transitions of the cesium atom.

Attitude

The orientation of a rigid body in space with respect to a given reference frame. The attitude can be expressed in various forms of ► *attitude parameters*.

Attitude Parameters

The set of variables used to parameterize the orientation of a rigid body in space. Their number ranges from three (e.g., ► *Euler angles*) to nine (e.g., ► *direction cosines*).

Automatic Dependent Surveillance (ADS)

A surveillance technique where each aircraft automatically broadcasts its own position periodically via data-link.

Automatic Direction Finding (ADF)

an electronic aid to navigation that identifies the relative bearing of an aircraft from a radio beacon transmitting in the medium or long-frequency bandwidth, such as a ► *Non Directional Beacon* or commercial radio broadcast station.

Availability

The probability that a user is able to determine its position with the specified accuracy and is able to monitor the integrity of its determined position at the initiation of the intended operation.

Azimuth

One of the coordinates in the ► *horizontal system*. Azimuth is the angle, measured positive towards the east, between north and the projection on the horizon of the direction in which an object is observed.

B

Bandwidth

The range of frequencies that pass through a system without (significant) attenuation.

Barycentric System

A coordinate system whose center (origin) is at the average center-of-mass of the solar system.

Baseband

The frequency range used by conventional transmitters or receivers for performing pre/post-processing of the desired information.

Baseband Signal

A signal with a near-zero frequency and low bandwidth, such as the GNSS spreading code and data modulated onto the carrier signal. Also termed the envelope signal.

Baseline

The separation between two points in 2-D or 3-D space. For ► *differential positioning* the baseline refers to the 3-D vector between two GNSS receivers, one set up on a ► *base station*, the other at a point (or in space) whose coordinate is to be determined relative to the ► *base station*. Baseline length is a scalar quantity, being the inter-receiver distance expressed in length units.

Base Station

► *Reference station*

BeiDou

A regional and global navigation satellite system implemented by China. The name refers to the constellation (Big Dipper or Plough) used to find the north direction in stellar navigation.

BeiDou Satellite-based Augmentation System (BDSBAS)

A ► *satellite-based augmentation system (SBAS)* being developed as part of the BeiDou Navigation Satellite System to provide horizontal and vertical navigation throughout China.

Best Linear Unbiased Estimator (BLUE)

The estimator with the best precision, i. e., smallest variance, of all linear unbiased estimators. Linear unbiased estimators are linear functions of the observables that have an expectation equal to the to-be-estimated unknown parameter vector.

Best Linear Unbiased Predictor (BLUP)

The predictor having the smallest mean square prediction error (best) of all linear unbiased predictors. Linear unbiased predictors are linear functions of the observables that have an expectation equal to the expectation of the to-be-predicted random parameter vector.

Between-receiver Difference

Difference between (code or carrier-phase) GNSS observations or parameters at the same frequency for two receivers tracking the same satellite.

Between-satellite Difference

Difference between (code or carrier-phase) GNSS observations or parameters at the same frequency of two satellites that are tracked by the same receiver.

Bias

In estimation, *bias* denotes the difference between the expected value of an estimator and the true value being estimated. Biases are often the consequence of an improper modeling of measurement processes, such as the neglect of non-negligible systematic errors. More loosely and generally, the term is often used to describe a systematic error or offset in a measurement.

Bias-to-noise Ratio (BNR)

A dimensionless measure of bias significance. The influential-BNR drives the probability of hazardous occurrence, while the testable-BNR drives the probability of missed detection.

Binary Offset Carrier (BOC) Modulation

An extension of ► *binary phase shift keying (BPSK) modulation*, in which the modulated signal is multiplied by an additional sine or a cosine square wave sub-carrier. Instead of a single lobe, the spectrum of a BOC-modulated signal exhibits two main lobes symmetrically shifted relative to the main carrier frequency. Therefore, BOC modulation is also known as a split-spectrum modulation. The separation of the two lobes is determined by the sub-carrier frequency. In GNSS, BOC signals are applied in order to fulfill spectral separation requirements between non-interoperable signals of different systems.

Binary Phase Shift Keying (BPSK) Modulation

A modulation scheme for radio navigation signals, in which the phase of the carrier is shifted by 0° or 180° depending on the binary value of the modulated signals (i. e., ranging code and navigation data).

Bit Error Correction

The process carried out within a GNSS receiver to identify and correct incorrectly received ► *navigation message* bits. This process requires the navigation message to contain redundant bits.

Bit Synchronization

The process carried out within a GNSS receiver to identify the epochs of navigation data bit/symbol transitions if the bit/symbol duration exceeds the primary code duration.

Blunder

A gross error in a measurement that is neither systematic nor of random nature.

Block

Different generations of Global Positioning System (GPS) satellites built by different manufacturers are commonly termed *blocks*. The second generation is further divided into Block II, IIA, IIR, IIR-M, and IIF satellites.

Boresight Angle

The angle between the line-of-sight direction and the symmetry axis of an antenna.

Boundary Layer

The lowermost atmospheric layer directly affected by the Earth's surface.

Box-wing Model

A simplified description of geometrical and optical (reflection/emission) properties of a GNSS satellite for the modeling of ► *radiation pressure* effects.

Broadcast Ephemeris

The ► *ephemeris* transmitted by a GNSS satellite as part of its ► *navigation message* to enable computation of the satellite positions within a receiver.

Broadcast Group Delay (BGD)

Alternative name for ► *Timing Group Delay (TGD)*.

C

C-band

The part of the spectrum of electromagnetic waves with carrier frequencies in the range from 4 GHz to 8 GHz

C/A-code

The ► *pseudo-random noise (PRN)* sequence used as coarse and acquisition ranging code within the Global Positioning System GPS. Each C/A-code has a length of 1023 chips and is transmitted in 1 ms. An entire family of C/A-codes has been defined for GPS as well as other radio navigation systems such as SBAS and QZSS. The serial number assigned to each code known as the PRN number and commonly used to identify the transmitting GPS satellite.

Cadastral Surveying

A form of land surveying for the determination, or marking-out of land property boundaries.

Carrier

A periodic electromagnetic wave on which the ranging code and navigation data of a radio navigation satellite system are modulated. GNSS signals are commonly located in the ► *L-band* and employ frequencies in a range of about 1100–1600 MHz.

Carrier Phase

The instantaneous phase (expressed in radians or cycles) of a periodic electromagnetic wave. In GNSS the term is commonly related to the beat phase of the received carrier after mixing with the nominal signal frequency, which represents a measure of the range variation between the transmitting satellite and the receiver.

Carrier Phase Ambiguity

The measurement of the ► *carrier phase* inside a GNSS receiver includes an arbitrary cycle count introducing an integer-cycle ambiguity. Depending on the specific tracking technique, half-cycle biases may also arise. The measured carrier phase range is furthermore affected by satellite or receiver specific biases causing a float-valued ambiguity in the measured carrier phase range.

Carrier-to-noise Density Ratio (C/N_0)

The ratio of the power level of the carrier signal to the noise power within a 1 Hz bandwidth.

Celestial Coordinates

Spherical coordinates of celestial objects with respect to a celestial reference system, called ► *right ascension* and ► *declination*, analogous to longitude and latitude, respectively.

Celestial Ephemeris Pole

The reference pole for ► *nutaton* and ► *polar motion* that was adopted by the 1980 IAU theory of nutation. By definition, it is a pole that exhibits no nearly diurnal motions in either the celestial or terrestrial reference frames.

Celestial Intermediate Pole

The reference pole for ► *nutaton* and ► *polar motion* that was adopted by the 2000 IAU theory of nutation.

It extends the definition of the ► *celestial ephemeris pole* by clarifying the distinction between nutation and polar motion. Motion of the celestial intermediate pole within the celestial reference frame that has a frequency between -0.5 cycles per sidereal day (cpsd) and +0.5 cpsd is defined to be nutation. Motion of the celestial intermediate pole outside this frequency band is defined to be polar motion. The polar motion parameters reported by Earth rotation measurement services give the location of the celestial intermediate pole within the rotating, body-fixed terrestrial reference frame.

Celestial Sphere

An imaginary sphere of infinite radius on which radial direction are projected. The center of the celestial sphere may be viewed as being geocentric or barycentric, depending on the definition of the associated coordinate system.

Central Synchronizer

The master clock of the GLONASS system. It comprises four ► *hydrogen maser* frequency standards and contributes to the mathematical GLONASS System Time steered to UTC(SU).

Certification

A process applied by a regulating body to determine whether an object has been manufactured according to an approved design and that the design ensures compliance with previously specified requirements. The requirements are often specified in ► *Minimum Operating Performance Standards (MOPS)* and related documents.

Cesium Beam Frequency Standard

A signal generator based on the hyperfine frequency of the cesium atom's ground state of 9 192 631 770 Hz. Typically, the cesium atoms are thermally formed into a beam and magnetically separated into the ground state for interrogation.

Chandler Period

A characteristic period in the oscillation of the Earth's rotation axis relative to its axis of figure discovered by S.C. Chandler. Due to the non-rigid structure of the Earth, the Chandler period may vary in a range of about 412–442 days.

Chapman Profile

An analytical model describing the height-dependent electron density variation in the ionosphere or its individual layers based on simplifying assumptions of the atmospheric structure as well as the ionization and recombination processes.

Chinese Area Positioning System (CAPS)

A positioning system developed in China, which transfers ground-generated navigation signals to users via communication satellites.

Choke-ring Antenna

An ► *antenna* composed of the sensitive antenna element and a surrounding, corrugated ground plane, which is typically made up of multiple concentric conductive rings. Choke ring antennas offer superior ► *multipath* suppression and are commonly used for

the highest accuracy user applications such as for geodetic surveys or ► *continuously operating reference stations*.

Circular Error Probable (CEP)

The radius of a circle centered on the true value that contains 50% of the actual measurements.

Clock Ensemble

A collection of clocks, not necessarily in the same physical location, operated together in a coordinated way to maximize the performance (time accuracy and frequency stability) and/or the availability of a time scale. Typically, the relative value of each clock is weighted, so that the best clocks contribute the most to the average.

Clock Offset

The offset between the reading of a satellite or receiver clock involved in the timing of GNSS measurements and a given reference time scale, such as the system time maintained by the GNSS control segment.

Coarse/Acquisition (C/A) Code

► *C/A-code*

Code Bias

A receiver and transmitter specific hardware ► *group delay* affecting the ► *pseudorange* (i. e., code observation) generation.

Code Division Multiple Access (CDMA)

A multiple access scheme that allows different transmitters to access the transmission channel simultaneously using the entire available bandwidth where the transmitter is identified by a unique code assigned to it. CDMA signals form the basis of most navigation satellite systems in use today.

Code Phase

The instantaneous phase of the ranging code modulated on a GNSS signal as sensed by a receiver. It can be expressed as the number of full and fractional code chips, or, alternatively as a time or distance value. The code phase is a measure of the transmission time (modulo the code duration) and used together with the current receiver time to form a ► *pseudorange* measurement.

Code Shift Keying (CSK)

A specific scheme used to increase the rate of navigation data in a GNSS signal. The ranging code is shifted relative to a nominal starting point by an amount that is determined by the encoded data word. Code shift keying is, for example, employed in the E6 signal of the Japanese ► *Quasi-Zenith Satellite System (QZSS)* to transmit high-rate correction data for ► *precise point positioning* to its users.

Coherent Integration Time

The time period chosen by a GNSS receiver to compute correlation values of the received signal with internal replica signals.

Cold Start

Activation of a GNSS receiver with no prior information on time, user position, and satellite orbit/clock data.

Compatibility

The ability of two navigation signals to be transmitted and used along each other without causing harmful interference for their users.

Conductive Test

Testing a receiver by directly connecting a signal source to the receiver.

Constrained Maximum Success-rate (CMS) Test

An ambiguity acceptance test which has the largest success rate for a given user-defined failure rate. This test requires the ► *integer least-squares (ILS)* solution as its input.

Construction Surveying

Land surveying that addresses the different positioning requirements of civil engineers and building professionals during the construction phase for any engineered structure.

Conterminous (or Contiguous) United States (CONUS)

The 48 adjoining US states and Washington, DC. CONUS is the portion of the United States that excludes Alaska, Hawaii, and all offshore territories.

Continental Drift

► *plate motion*

Continuity

The probability that a user is able to determine its position with the specified accuracy and is able to monitor the integrity of its determined position over the time interval applicable for the corresponding phase of flight. Assuming the service is available at the start of an operation, this is the probability of it becoming unavailable over a specified time interval linked to the duration of the operation.

Continuously Operating Reference Stations (CORS)

A GNSS receiver and antenna established on a permanent and stable site that serves as a ► *control point* of a geodetic network or reference station for ► *differential GNSS* systems.

Controlled Flight into Terrain (CFIT)

The situation when a normally functioning aircraft under the complete control of the pilot is inadvertently flown into an obstacle.

Control Point

A marker or monument used for surveying, typically with known coordinates in the local or national geodetic ► *datum*. May also be a ► *base station* if a GNSS receiver is operated at that point.

Control Segment

The ground infrastructure used to operate a global or regional navigation satellite system.

Coordinated Universal Time

► *Universal Time Coordinated (UTC)*

Coordinate Time (TCB, TCG)

A set of fundamental relativistic time scales with rates based on the SI second in the respective reference frames, i. e., at the Solar System's center of gravity for Barycentric Coordinate Time (TCB) and the Earth's center for Terrestrial Coordinate Time (TCG).

Correlation

A measure of agreement between two statistical values or time series. In estimation, correlation is

defined as a normalized form of the ► *covariance*. In GNSS signal processing, the correlation describes how well two signals or ► *pseudo-random noise* codes match each other.

Correlator

A device used inside a GNSS receiver to measure the ► *correlation* of the incoming signal and a receiver-generated replica. The correlator values serve as input for the ► *tracking loops*, which aim to continuously align the replica code and phase with the incoming signal. In order to best measure the instantaneous code offset, a combination of an *early* and a *late* correlator is used, which employ time shifted code replicas. The early-minus-late difference of the correlator outputs can then be used as a ► *discriminator* to sense the tracking error.

Correlator Spacing

The spacing (in units of PRN chips) between the early and late ► *correlators* in a conventional ► *delay lock loop*.

Coseismic

means *during an earthquake*. The term is most commonly used to describe displacements that result from an earthquake.

COSPAS-SARSAT

An international satellite-based search and rescue system.

Costas Loop

A special form of a ► *phase lock loop* developed by J.P. Costas that uses a two-quadrant phase discriminator to track a signal with ► *binary phase shift keying (BPSK) modulation* without being affected by data bit transitions.

Covariance

A measure of how much two random variables change together. ► *Correlation* is the normalized version of covariance. It is obtained by dividing the covariance by the standard deviations of the two random variables.

Cramer Rao Lower Bound (CRLB)

A lower bound on the variance of unbiased estimators of deterministic parameters. Named after H. Cramer and C.R. Rao

Cross-correlation

Operation on two (real or complex) Doppler and delay-aligned signals to estimate their temporal coherence, as a function of the relative delay. Such function is termed waveform.

Cycle Slip

An unknown discontinuity in the measured ► *carrier phase* usually resulting from a temporary loss-of-lock (e.g., due to shading) in the carrier tracking loop of a GNSS receiver.

D

Data Bit

The basic information unit of the navigation message modulated on a GNSS signal. This unit is called

symbol, if the navigation message employs a ► *forward error correction* scheme. Otherwise it is called bit.

Data Channel

A GNSS signal component used to broadcast navigation data. For improved measurement performance, the data channel is complemented by a ► *pilot signal* in modern GNSS signals.

Data Demodulation

The process carried out within a GNSS receiver to extract a data bit or data symbol from a received GNSS signal.

Data Snooping

A procedure to identify the observations contaminated with gross errors.

Data Symbol

► *data bit*

Datum

A set of parameters and conventions that defines and realizes a coordinate system for geodetic control on a national or global scale. Nowadays realized by the 3-D Cartesian coordinates or 2-D geodetic coordinates of a network of ► *control points*.

Decision Altitude or Height (DA/H)

The altitude or height during an instrument approach procedure where the pilot must decide to either continue the approach to land or initiate the missed approach procedure. The decision is based on the availability of the required visual references. The decision altitude is measured above mean sea level and a decision height is above the ground. The DA/H is the point at which a missed approach is initiated and does not preclude the aircraft from descending below this height before the aircraft starts to climb.

Declination

The angle, at right-angles to the ► *equator*, measured between the equator and a celestial body. Together with ► *right ascension*, declination forms the equatorial system of coordinates.

Deflection of the Vertical

The angle between the tangent to the plumb line (direction of gravity) and the normal to the ellipsoid at a point.

Deformation

Any displacement that changes the shape of a body, as opposed to rigid body motion. Active tectonic and volcanic processes cause both recoverable and permanent deformation of the Earth.

Delay Lock Loop (DLL)

A controller used within a GNSS receiver to align a replica of the ranging code with the incoming signal after removal of the carrier. The DLL combines a steerable code generator with a code ► *discriminator* to sense the tracking error and a loop filter to reduce the tracking noise.

Differential Code Bias (DCB)

Difference of either receiver or satellite hardware biases on the code (or ► *pseudorange*) observations of two frequencies.

Differential GNSS (DGNSS)

GNSS positioning technique based on the principle that common biases between receivers can be eliminated or significantly reduced by processing multi-receiver code (pseudorange) and/or carrier-phase data simultaneously tracked from the same satellites. At least two receivers are needed; one is the reference and the other rover. This technique can be implemented either by differencing the observations of reference and rover, or by transmitting corrections determined at the reference to the rover.

Dilution of Precision (DOP)

A scalar measure that captures the impact of the instantaneous receiver-satellite geometry on the precision of ► *single point positioning*. It is computed as the square root of the sum of those diagonal elements of the variance matrix excluding the variance factor for which the DOP needs to be evaluated (e.g., Position DOP based on the diagonal elements for east, north and up and horizontal DOP when only the diagonal elements for east and north are involved, etc.).

Direct Conversion

A method to digitize a radio frequency signal without prior down-conversion.

Direction Cosine

the cosine of the angle formed by a vector in space and a reference direction.

Discriminator

A function of ► *correlator* values used to measure the code, phase, or frequency offset between the incoming GNSS signal and the replica generated in the receiver.

Dispersive Medium

A medium in which an electromagnetic wave propagates for which the magnetic constant (permeability) and/or the dielectric constant (permittivity) depend on frequency. Therefore, the phase velocity as well as the group velocity of an electromagnetic wave in a dispersive medium also depend on frequency.

Displacement

The change in position of a point. All points on the Earth's surface are displaced slowly and steadily by plate motions, but sudden displacements also can occur due to events such as earthquakes. Seasonal movements of mass cause quasi-periodic displacements due to loading.

Disposal Orbit

An orbit for satellites that are beyond the end of their service life, which is designed to minimize the probability of a collision with any other satellites or space debris.

Distance Measuring Equipment (DME)

A combination of ground and airborne equipment which provides a continuous slant range distance-from-a ground station by measuring the propagation delay of a signal transmitted by the aircraft to the station and responded back. The ground equipment is a VHF transmitter and receiver (called the transponder) and the airborne equipment is called the interrogator.

Doppler/Delay Alignment

Operation to compensate the delay experienced by a signal due to the relative motion of the transmitter and the receiver. For a short time interval the relative motion can be modeled using two quantities, the initial relative delay and the initial relative Doppler.

Doppler Delay Map (DDM)

An evaluation of the ► *cross-correlation* function of the received signal with a replica signal expressed as a function of code delay and ► *Doppler shift*.

Doppler Effect

► *Doppler Shift*.

Doppler Range

The range of frequency offsets from the nominal signal frequency considered by a GNSS receiver in the signal search to account for the Doppler shift due to the relative motion of transmitter and receiver as well as the frequency error of the local oscillator.

Doppler Shift

The frequency shift experienced by an electromagnetic (or acoustic) wave due the relative motion of the transmitter and receiver. Named after the Austrian nineteenth-century physicist Christian Doppler.

Double-difference

Difference between either two ► *between-receiver differenced* observations/parameters that correspond to different satellites, or two ► *between-satellite differenced* observations/parameters that correspond to different receivers.

Down-conversion

A method to convert the frequency of a radio-frequency signal to a reduced intermediate frequency by mixing it with a harmonic signal. Down-conversion facilitates the analog-to-digital conversion and the subsequent signal processing.

Draconitic Period

The time that elapses between two passages of the orbiting object through its ► *ascending node*.

Draconitic Year

The repeat period of the orientation of a GNSS constellation with respect to the Sun, e.g., 351.5 days for GPS.

Dual-frequency

The use of GNSS measurements on two different signal frequencies, e.g., to eliminate the impact of ionospheric path delays.

Dynamical Time (TDB, TDT)

A relativistic time scale having a rate that matches the SI second at the Earth's surface, defined as a re-scaling of the ► *Coordinate Time scales TCG and TCB*. In 1991 it was agreed by the International Astronomical Union that Terrestrial Dynamical Time (TDT) should just be called ► *Terrestrial Time (TT)*.

E**Early-minus-late Correlator**

► *correlator*

Earth-centered Earth-fixed (ECEF)

Refers to a coordinate system or frame that is fixed to the Earth's crust with a conventional orientation and whose origin is at the Earth's center of mass. The axes of the ► *International Terrestrial Reference Frame* provide a possible choice of an ECEF coordinate system.

Earth-centered Inertial (ECI)

A non-accelerating, non-rotating frame aligned with the ► *Earth-centered Earth-fixed (ECEF)* frame at a specific instant of time. Various approximate ECI frames may be convenient for analysis. For example, the non-rotating ECI frame origin may be assumed to coincide with the ECEF origin, which is approximate because the ECEF origin is accelerating.

Earth Model PZ-90

► *Parametry Zemli 1990 (PZ-90)*

Earth Oblateness

The oblateness or ► *flattening* is a measure of how much the Earth's elliptical shape differs from a sphere.

Earth Orientation Parameters

A set of parameters that relate an Earth-centered Inertial coordinate system to an Earth-centered Earth-fixed coordinate system and vice versa. It consists of the small angles that define the motion of the ► *Celestial Intermediate Pole* (approximately Earth's instantaneous spin axis) with respect to the terrestrial reference system (► *polar motion*) and its motion relative to the celestial reference system (► *precession* and ► *nutation*).

Earthquake Cycle

The roughly cyclic buildup of stress and strain and their release in earthquakes is termed the earthquake cycle. The shallow part of most faults is stuck together (locked) by friction most of the time, while the deeper part continues to creep at a nearly steady rate. This causes an increase in the stress around the locked fault zone and results in elastic strain energy being stored in the crust. When the driving stress exceeds the frictional resistance, the fault slips and an earthquake occurs. Earthquakes are never perfectly periodic in occurrence, so this should not be understood to be a periodic process.

Earth Radiation Pressure

The non-gravitational force acting on a satellites due to radiation reflected/emitted by the Earth's surface (► *Albedo*).

Earth Rotation Angle

The angle between the origins on the intermediate terrestrial and celestial equators; it is proportional to the time associated with Earth's rate of rotation.

Ecliptic

The imaginary great circle representing the intersection of the plane of the Earth's orbit with the celestial sphere.

Effective Isotropic Radiated Power (EIRP)

The amount of power radiated by an antenna in a particular direction as referenced to the amount of power fed to an idealized lossless isotropic antenna that would produce the same power density.

Elevation

The angle measured from the observer's horizontal plane with respect to the station-satellite line-of-sight. May also refer to a station's height in metric units with respect to the zero datum level, or synonymous with ► *ellipsoidal height*.

Elevation Mask

A threshold controlling the allocation of GNSS for tracking in the receiver or for processing in the positioning. Only satellites exceeding the defined minimum angle above the ► *horizon* (or alternatively the antenna ground plane) will be considered.

Ellipsoid

In geodesy, it generally refers to a geometric figure defined by rotating an ellipse about its minor axis and whose parameters (size and flattening) are chosen to yield a good approximation to the geoid. It is the mapping surface for horizontal geodetic control. It is also called a spheroid to distinguish it from a more general tri-axial ellipsoid.

Ellipsoidal Height

The distance along the perpendicular (normal) to an ellipsoid starting from the ellipsoid. Ellipsoidal height is positive for points outside the ellipsoid and negative for points inside the ellipsoid, being zero on the ellipsoid.

Elongation

The angle between two bodies, as seen by an observer.

Empirical CODE Orbit Model (ECOM)

A model for solar ► *radiation pressure* developed at the Astronomical Institute of the University Bern, a member of the Center for Orbit Determination in Europe (CODE). The radiation pressure is modeled with constant and sine/cosine terms in a Sun-oriented system.

Ephemeris

Based on a Greek expression (*for a single day*), the term is widely used for astronomical tables giving the daily coordinates of a planet or other solar system body. In GNSS, ephemerides are likewise tabular positions and clock offsets of the GNSS satellites (resulting, e.g., from a ► *precise orbit determination*). Furthermore, the term ephemeris is used for a set of orbital parameters transmitted by a GNSS satellite as part of its ► *navigation message* to enable computation of the satellite positions within a receiver.

Ephemeris Time (ET)

An astronomical time defined by the orbital motions of the Earth, Moon, and planets. Ephemeris Time was introduced in 1952 to be provide a time scale independent of the irregular, unpredictable variations in the rotation of the Earth, inherent to the ► *Universal Time* in use beforehand. Within the framework of relativistic theories of motion, Ephemeris Time has been replaced by ► *Terrestrial Time (TT)*.

Equator

An imaginary great circle on the celestial sphere, which is perpendicular to the Earth's axis of rotation. The equator separates the northern and southern

celestial hemispheres, and is simultaneously the reference plane for the equatorial system of coordinates, which uses the coordinates ► *right ascension* and ► *declination*. It defines the plane that is orthogonal to the polar axis of a global coordinate system. For the Earth it is the circle on the ellipsoid at zero latitude.

Equatorial Coordinates

Coordinates referred to the ► *equator* (► *right ascension*, ► *declination*).

Equinox

► *vernal equinox*.

Euler Angles

A set of three angles that define the orientation of a rigid body. The Euler angles define a sequence of three consecutive rotations that enable aligning any two arbitrarily rotated orthogonal frames.

Euler Axis and Angle

The rotation axis and rotation angle of a body in space relative to an initial or reference orientation. Any rotation of a rigid body in space can be described in terms of an axis-angle parameterization.

European Geostationary Navigation Overlay Service (EGNOS)

A ► *satellite-based augmentation system (SBAS)* developed by the European Space Agency, the European Commission, and EUROCONTROL to provide horizontal and vertical navigation throughout Europe. It has provided safety-of-life service since 2011.

Expandable Slot

In the GPS constellation, one of three orbital positions that can each be divided into a pair of positions when the constellation has more than the nominal number (24) of satellites.

Extended Kalman Filter (EKF)

The ► *Kalman filter* applied to the linearized versions of non-linear state space measurement and dynamic models.

F

Fading Frequency

The phase rate-of-change (known simply as phase rate) of a multipath signal relative to the direct-path signal.

Fixed Solution

► *ambiguity fixed solution*

Fixed Failure Rate Ratio Test

An ambiguity acceptance test that is computed as a ► *ratio test*, but that has a guaranteed failure rate. The required failure rate can be set by the user.

Flächenkorrekturparameter (FKP)

Original German designation of area correction parameters (horizontal gradients) of regional models for distance-dependent biases transmitted to the users as part of ► *network RTK corrections*.

Flattening

The geometric parameter of an oblate ► *ellipsoid* defined by the ratio of the difference between

semi-major and semi-minor axes of an ellipsoid to its semi-major axis.

Flicker Noise

A type of noise in electronic systems (e.g., oscillators), which exhibits a power spectral density inversely proportional to the frequency.

Flight Management System (FMS)

An aircraft computer system with multiple functions for managing a flight. The FMS includes navigation and guidance functions and contains a database allowing flight plans and routes to be pre-programmed.

Flight Technical Error (FTE)

The difference between the estimated position and the defined path. It serves as a measure of how well the pilot or the avionics can follow the guidance information provided by the navigation system.

Float Solution

► *ambiguity float solution*.

Footprint

The region beneath a satellite that can receive and utilize its signal. A navigational satellite footprint is often depicted as a circular region directly below the satellite representing a set of users whose line of sight to the satellite is at least 5° above their local horizon.

Forecast

A forecast describes the future state and/or development of characteristic system parameters like the temperature in the troposphere or the electron density in the ionosphere based on the current state and additional information.

Forward Error Correction (FEC)

A technique used to minimize bit errors in data transmission which introduces redundant information into the data stream. Other than simple parity checksums, FEC enables not only detection but also correction of errors.

Frame Bias

A small constant angular offset between the current kinematic ICRF pole and origin (in right ascension) and the dynamic pole and equinox of J2000.

Frame synchronization

The process carried out within a GNSS receiver to identify the beginning of the navigation data message.

Free-space Loss

The change in signal power of an electromagnetic wave emerging from an isotropic radiator when propagating in free space. The free-space loss grows with the square of the distance and is inversely proportional to the square of the wavelength.

Frequency

The rate of a repetitive event, i.e., the inverse of its repeat period. In the SI system of units the period is expressed in seconds (s), and the frequency is expressed in Hertz (Hz).

Frequency Lock Loop (FLL)

A controller used to align the frequency of a carrier replica inside a GNSS receiver with that of the incoming signal. It comprises a ► *numerically controlled oscillator (NCO)*, a frequency

► *discriminator* that senses the instantaneous tracking error and a loop filter that provides a smoothed estimate of the frequency error for feedback to the NCO.

Frequency Division Multiple Access (FDMA)

A multiple access scheme that allows different transmitters to access the channel simultaneously but using slightly different frequencies within the specified overall bandwidth. For global satellite navigation systems, FDMA signals are presently only used by ► *GLONASS*

Friis Formula

A formula named after the electrical engineer H.T. Friis, which is widely used in communications to calculate the total noise factor of a cascaded stage radio frequency ► *frontend* of which each stage is represented by a noise factor and a gain.

Frontend

The combination of different modules that convert the incoming ► *baseband signal*, on the transmitter side, to the specified radio frequency before passing it on to the antenna. A receiver frontend does the opposite.

G

Galileo

The European global navigation satellite system.

Galileo-GPS Time Offset (GGTO)

As GPS and Galileo use different reference time systems, there is a system time offset between the two systems, the Galileo-GPS time offset. This offset can be several tens of nanoseconds or tens of meters. Users can correct their data for the offset, since the GGTO is broadcast as part of the navigation messages.

Geocentric Coordinates

Coordinates referred to the center of the Earth.

Geodetic Coordinates

Coordinates of latitude, longitude, and height associated with a particular ► *ellipsoid*. The geodetic latitude is the angle of the ellipsoid normal for a point relative to the Equator. The geodetic longitude is the same as the spherical longitude. See also ► *ellipsoidal height*.

Geodetic-grade Receiver

Top-of-the-line GNSS receiver able to make ► *carrier phase* and ► *pseudorange* measurements on multiple L-band frequencies broadcast by several GNSS constellations.

Geodetic Reference System 1980 (GRS80)

The current internationally defined reference ellipsoid and associated gravitational field for geodetic applications.

Geographic Coordinates

A general name that refers to coordinates associated with the spherical shape of the Earth

Geographic Information System (GIS)

A software system that manages, analyses, displays spatial data that has been organized in *layers*. For the purpose of creating special feature maps, undertaking spatial analysis, assist in decision-making, and more.

Sometimes the term may also refer to the spatial datasets themselves.

Geoid

A surface on which the Earth's gravity potential is a constant (equipotential or level surface) and that closely approximates global mean sea level.

Geoid Height

The ► *ellipsoidal height* of the ► *geoid*; also known as the geoid undulation.

Geoid Model

Representation of the ► *geoid height* across a local area, a region, or the globe. May be in the form of geoid height contours on the ► *reference ellipsoid*, as point values, or gridded data, or various mathematical functions – the most common being in the form of spherical harmonics.

Geoid Undulation

Same as ► *geoid height*.

Geostationary Earth Orbit (GEO)

A circular orbit at approximately 36 000 km altitude above the Earth's equator with an orbital period equal to the Earth's rotational period. A satellite in a geostationary Earth orbit will appear to be in a fixed position in the sky for terrestrial observers.

Geostationary Satellite

A satellite in a ► *geostationary Earth orbit*. Often used for communication and for ► *satellite-based augmentation systems*.

Global Navigation Satellite System (GNSS)

A navigation system that can provide a positioning solution anywhere in the world. The term is currently used collectively for GPS, Galileo, GLONASS, and BeiDou.

Global'naya Navigatsionnaya Sputnikova Sistema (GLONASS)

Global navigation satellite system of the Russian Federation, providing global permanent positioning, navigation and timing service for land, air and space users worldwide. GLONASS is a dual-use system providing both authorized and open access services.

Global Positioning System (GPS)

A global navigation satellite system owned and operated by the United States Air Force.

Gold Code

A family of ► *pseudo-random noise (PRN)* sequences with good auto and cross-correlation properties proposed by Robert Gold. Gold codes are, for example, employed for the GPS coarse and acquisition (► *C/A*) code.

GPS Aided GEO Augmented Navigation (GAGAN)

A ► *satellite-based augmentation system (SBAS)* developed by the Indian Space Research Organization, the Airports Authority of India, and the Directorate General of Civil Aviation to provide horizontal and vertical navigation throughout India. It has provided safety-of-life service since 2014.

GPS Time

The timescale used by GPS, which began at midnight ► *Universal Time Coordinated (UTC)* on January 5/6, 1980 and is not adjusted by leap seconds.

GPS Week

The integer number of weeks elapsed since the start of ► *GPS time*. GPS weeks start at midnight from Saturday (day-of-week 6) to Sunday (day-of-week 0). In accordance with the 10-bit representation of the GPS week in the legacy GPS ► *navigation message*, its value is often given modulo 1024, resulting in roll-overs in August 22, 1999 and April, 7 2019. RBL,OM)

Greenwich Mean Time (GMT)

A 24-hour time keeping system whose hours, minutes, and seconds represent the time-of-day at the Earth's prime meridian (0° longitude) located near Greenwich, England. GMT was adopted as the world's first global time standard in 1884. GMT no longer exists, since it was replaced by other astronomical time scales many years ago, which in turn were subsequently replaced by the atomic time scale UTC.

Greenwich Hour Angle

The angle at a specified epoch between the *x*-axes of the ► *Earth-centered inertial* and ► *coordinate systems*.

Grid Ionospheric Vertical Error (GIVE)

A parameter broadcast by SBAS to indicate the possible magnitude of the vertical delay error for a specific ionospheric grid point delay estimate. GIVE is determined from a broadcast 4-bit number, called the GIVE indicator or GIVEI. A look up table is used to convert the indicator to a 1-sigma overbound value (► *overbound*) called the σ_{GIVE} . By tradition, GIVE itself is a 99.9% number or $3.29 \times \sigma_{\text{GIVE}}$. A relevant set of σ_{GIVE} s is used to interpolate the corresponding User Ionospheric Vertical Error overbound, σ_{UIVE} . This vertical overbound is finally multiplied by the obliquity factor to obtain the complete User Ionospheric Range Error overbound, σ_{UIRE} .

Ground-based Augmentation System (GBAS)

A local area differential GNSS augmentation system using multiple airport-based reference receivers that provide corrections and integrity information to users via a very-high frequency (VHF) data link known as the ► *VHF data broadcast (VDB)*. GBAS supports aircraft performing precision approach and landing operations as well as other operations near GBAS-equipped airports.

Ground Plane

An electrically conducting flat surface at the *bottom* of an ► *antenna* that reflects electromagnetic waves. Optimum antenna properties (such as ► *multipath resistance*) are obtained if the dimension of the ground plane is large compared to the ► *wavelength*.

Ground Segment

A major component of a ► *global navigation satellite system* providing satellite control and constellation keeping, as well as orbit and clock data calculation for mission operations.

Ground Station

A terrestrial radio station enabling communication with a satellite. Depending on the specific application, one may distinguish between various functions:

telemetry stations for reception of satellite data on the ground, telecommand (or uplink) stations for sending control commands to the spacecraft, or tracking stations providing range, range rate or angular measurements for ► *orbit determination*.

Group Delay

is a measure of the propagation time of the amplitude envelope of an electromagnetic signal through a device or medium. In the context of GNSS, it describes the delay experienced by the code modulation upon signal propagation through a dispersive medium such as the ionosphere or parts of the signal generating, transmitting, or receiving equipment.

Group Velocity

Speed of propagation of the envelope, i. e., the code signal of navigation signals, of a modulated electromagnetic wave. It is a measure for the speed of movement of the wave energy.

H**Hand-Over Word (HOW)**

The second 30-bit word within every GPS legacy navigation data 300-bit sub-frame. HOW includes a time stamp.

Harmonic

A signal whose frequency is an integer multiple of some other signal's frequency. Nonlinearity in any one of several stages involved in radio frequency transmission generates (typically undesired) power at harmonics of the transmission frequency.

Hatanaka Compression

A loss less technique used to compress GNSS data files in ► *Receiver Independent Exchange (RINEX)* format.

Helmert Transformation

A 7-parameter similarity transformation named after the geodesist and mathematician F. R. Helmert. It relates two frames through a shift vector, a rotation and a scale factor.

Higher-order Ionospheric (HOI) Terms

Contributions to the ionospheric delay of a GNSS signal, which depend on higher than second-order terms of the frequency and cannot be eliminated by the ► *ionosphere-free combinations* of two observations.

Horizon

The imaginary line of intersection between a plane tangent to the surface of the Earth at the observer and the celestial sphere. The horizon is the reference plane for the ► *horizontal system* with coordinates ► *azimuth* and ► *altitude*.

Horizontal System

A coordinate system related to the local horizon of an observer, and where the coordinates used are ► *azimuth* and ► *elevation*.

Hot Start

Activation of a GNSS receiver with prior information on the approximate time and user position as well as

the broadcast ephemeris of the GNSS satellites to speed up the signal search and acquisition and to enable an immediate computation of the navigation solution after collection of valid ► *pseudorange* measurements.

Hour Angle

Difference between the local ► *sidereal time* and the ► *right ascension* of a star. The hour angle measures the sidereal time that has passed since the last culmination.

Hydrogen Maser

A signal generator based on the hyperfine transition of the neutral hydrogen atom at 1420.405752 MHz.

Hydrographic Surveying

A form of surveying in support of offshore engineering (such as associated with pipelines, undersea cables, breakwaters, harbor works, dredging, etc.) and sea floor charting.

Hydrostatic Delay

The dry component of the ► *slant total delay*.

Hypothesis Testing

is a formalized decision rule of rejecting or not rejecting the null hypothesis. The null hypothesis, representing the model one believes to be true, is thereby compared to one or more alternative hypotheses.

I

IF-level Simulator

A simulator that generates intermediate frequency (IF) samples of GNSS signals similar to those expected at the output of the receivers' front end. These samples can be processed directly by the digital processor of a receiver, or, with a digital-to-analog converter and an up-converter, can be converted to RF signals for conductive or rebroadcast testing.

Inclined Geosynchronous Orbit (IGSO)

An orbit synchronous with the Earth's rotation period of about 24 h but inclined with respect to the Earth's equator. IGSO satellites are commonly used for regional navigation satellite systems as a complement to strictly ► *geostationary satellites*.

Inertial Measurement Unit

An electronic measurement device combining accelerometers and gyros to measure the specific force and angular rate of the body to which the sensor is attached.

Inertial System/Frame

A non-rotating system or frame in free fall with respect to the ambient gravitational field. *Pseudo* or *quasi* are sometimes appended to distinguish it from the original Newtonian concept of a frame at rest or having only constant rectilinear motion.

Influential Bias

A ► *bias* that propagates into the parameter estimator; such bias lies in the range space of the design matrix.

A non-testable bias is always influential.

In-phase (I) Component

GNSS signals frequently use ► *quadrature phase shift keying* to transmit two orthogonal signal

components, such as a civil and a military signal, on one frequency. In the case of GPS L5, the in-phase (I) component transmits the navigation data (► *data channel*), while the quadrature (Q) component (which is 90° out of phase) carries a dataless ► *pilot signal*.

Integer Ambiguity

The unknown number of wavelengths (cycles) that is contained in the measured carrier phase range from the receiver to satellite antenna.

Integer Ambiguity Resolution

The procedure by which the unknown double-differenced carrier-phase ambiguities are estimated and validated as integers. Once these ambiguities can be considered known, the corresponding carrier-phase measurements will act as very precise pseudorange measurements.

Integer Bootstrapping (IB)

Integer estimation that is based on a combination of ► *integer rounding* and sequential conditional least-squares estimation. The success rate of integer bootstrapping is never smaller than that of integer rounding and never larger than that of ► *integer least-squares*.

Integer Least-squares (ILS)

Integer estimation that is based on the principle of least-squares. The ILS estimator has the largest success rate of all integer estimators. Its success rate is invariant for ► *Z-transformations*. In contrast to ► *integer rounding* and ► *integer bootstrapping*, the ILS-estimation requires an integer search.

Integer Rounding (IR)

Integer estimation that is based on scalar rounding to the nearest integer. The success rate of integer rounding is never larger than that of ► *integer bootstrapping* or ► *integer least-squares*.

Integrity

A measure of the trust that can be placed in the correctness of the position solution. Integrity includes the ability of a system to provide timely and valid warnings (alerts) to the user.

Inter-channel Bias (ICB)

Difference in code or phase biases of GLONASS ► *FDMA* signals transmitted on different frequency channels.

Interface Control Document (ICD)

A document describing the interfaces between the GNSS ► *space segment* and the ► *user segment*. It provides a description of the signal structure and modulation, the format and contents of the navigation data, as well as relevant algorithms for using these data in the positioning.

Interference

Radio frequency power in one or more GNSS bands that degrades a GNSS receiver's ability to acquire and track GNSS signals. Interference may be structured, as in GNSS ► *spoofing*, or unstructured, as in wideband Gaussian noise jamming. It may be intentional, as in deliberate jamming, or unintentional, as in noise generated by electronics surrounding a GNSS receiver.

Inter-frequency Bias (IFB)

Difference in code or phase hardware biases of signals with different frequencies.

Intermediate Frequency (IF)

A frequency lower than the carrier frequency to which the modulated GNSS signal is shifted to facilitate processing in the signal generation or processing chain.

International Atomic Time (TAI)

A time scale realized by a weighted average of the time kept by over 400 atomic clocks worldwide. It is the basis for ► *Coordinated Universal Time (UTC)*, which is used for civil timekeeping all over the Earth's surface. TAI is computed monthly by the International Bureau of Weights and Measurements (BIPM).

International Celestial Reference System (ICRS)

A system of coordinates and conventions that define coordinates of objects in inertial space. It is defined and maintained by the International Earth Rotation and Reference Systems Service (IERS).

International Celestial Reference Frame (ICRF)

A reference frame that realizes the ► *International Celestial Reference System* by means of coordinates of objects accessible directly by radio-astronomical observations.

International Terrestrial Reference System (ITRS)

A system of coordinates and conventions that define coordinates of points tied to the Earth. It is defined and maintained by the International Earth Rotation and Reference Systems Service (IERS).

International Terrestrial Reference Frame (ITRF)

A reference frame that realizes the ► *International Terrestrial Reference System* at a particular epoch by means of coordinates of definite points that are accessible directly by occupation or by observation. The IERS is responsible for computing the coordinates of this global network in order to realize the ITRF.

Interoperability

The ability of jointly using two independent navigation systems for with a resulting benefit over the use of each individual system. ► *Compatibility* of their signals is a prerequisite for the interoperability of two systems.

Interplex

A special case of a phase-shift-keyed/phase-modulated (PSK/PM) multichannel system that is characterized by high power efficiency for a number of components.

Inter-seismic

Between earthquakes. The inter-seismic period is the time period between major earthquakes on a given fault. In most models of the earthquake cycle, the rate of deformation is expected to be constant or varying only slowly for most of the inter-seismic period.

Inter-system Bias (ISB)

Difference in receiver hardware biases between signals of two GNSS constellations plus the offset between the time scales of the two systems.

Inter-satellite Type Bias (ISTB)

Difference in receiver hardware biases between signals transmitted by different satellite types within a constellation. In the case of BDS, the existence of inter-satellite type biases has been demonstrated for signals of geostationary satellites that are combined with those of other (IGSO, MEO) satellites.

Inter-Signal Correction (ISC)

Correction values in the GPS modernized navigation message to enable a consistent processing of different navigation signals using a single set of clock offset values. ISCs represent a specific form of ► *differential code biases*.

Ionosphere

The ionized part of the upper atmosphere from 50 km up to about 1000 km height. The ionospheric plasma is primarily produced by electromagnetic and particle radiation transmitted from the Sun. Consequently, the ionization level depends strongly on geographic location, season and local time. Spatial structure and dynamics of the ionospheric plasma are strongly coupled with other geospheres such as atmosphere and magnetosphere causing a high variability of the plasma density.

Ionosphere-free Combination

A linear combination of two GNSS observations, which eliminates the dominant contributions of ionospheric path delays that vary with the inverse square of the signal frequency.

Ionospheric Correction Models

In a first-order approach the ionospheric delay or range error is proportional to the ► *total electron content (TEC)* along the ray path. Therefore, to reduce ionospheric range errors in single frequency GNSS applications, ionospheric electron density models (► *NeQuick* for Galileo) or simple TEC models (e.g., ► *Klobuchar* model for GPS or NTCM) can be used.

Ionospheric Gradient

Steep variations in time and space of the ionosphere electron content, which cause large delays in the GNSS differential observations even for small baselines.

Ionospheric Grid Point (IGP)

A specific reference location on an imaginary thin shell above the Earth's surface. In ► *satellite-based augmentation systems (SBAS)* a set of IGPs is used to describe the ionosphere by transmitting ionospheric delay values for a fixed set of locations over the region of interest.

Ionospheric Perturbations

There exist several types of ► *space weather* driven temporal and spatial electron density perturbations in the ionosphere and plasma sphere that may impact GNSS in different ways. Besides radio ► *scintillations* caused by small scale electron density irregularities also medium scale and large scale ► *traveling ionospheric disturbances* (MSTIDs, LSTIDs) or solar flare induced sudden ionospheric disturbances (SIDs) may degrade the GNSS

performance in precision and safety of live applications.

Ionospheric Pierce Point (IPP)

A location, often specified by latitude and longitude, between the user and a satellite where the line of sight between the two intersects an imaginary thin shell above the Earth's surface, a shell in which all electrons are assumed to be concentrated. Most commonly the thin shell is specified to have a constant height of 350 km above the reference ► *ellipsoid* of the 1984 release of the ► *World Geodetic System (WGS-84)*

Ionospheric Refraction

The signal propagation delay and signal path bending induced by free electrons in the upper part of the atmosphere. Besides the plasma density along the ray path, ionospheric refraction depends on the radio wave frequency (dispersion) and the geomagnetic field vector along the ray path (anisotropy).

J

J2000

A standard epoch representing midday on January 1, 2000 (2000 Jan. 1.5 = JD 2451545.0).

Julian Day

A time unit of ► *dynamic time*. It is exactly 1/36525 of a Julian century.

Julian Day (JD) Number

An integer day number obtained by counting days from the starting point of noon on January 1, 4713 BC (Julian Day zero).

K

Kalman Filter

A recursive algorithm to obtain a ► *minimum mean squared error (MMSE) estimate* of the state vector of a linear dynamic system based on a time series of past and current observations. Each cycle of the recursion consists of a time-update and a measurement update. In the time-update the dynamic model is used to predict the state vector from the previous epoch, while in the measurement-update the measurements of the current epoch are used to improve upon the estimate of the predicted state vector.

Keplerian Elements

A set of six orbital elements used to describe the shape and spatial orientation of the orbit, typically comprising the semi-major axis, eccentricity, inclination, longitude of the ascending node, argument of perigee, and mean anomaly.

Keplerian Orbit

The trajectory of a satellite around a central point mass, for the special case of an attracting force proportional to the inverse square of the distance. Bound Keplerian orbits are ellipses confined to a fixed ► *orbital plane*.

Klobuchar Model

An empirical model used to describe the ionospheric time delay in single-frequency GNSS measurements. The Klobuchar ionospheric model was first adopted by the Global Positioning System but is also used by various other GNSSs in the same or slightly modified form. It comprises eight parameters that are broadcast by the GNSS satellites with at least daily updates.

Kinematic Positioning

Estimation of epoch-wise coordinates for a moving or non-moving GNSS receiver from observations covering a given data arc.

Korean Augmentation Satellite System (KASS)

A ► *satellite-based augmentation system (SBAS)* being developed by the Ministry of Land, Infrastructure, and Transport to provide horizontal and vertical navigation throughout the Korean peninsula.

L

Laser Time Transfer (LTT)

A technique for synchronization of space and ground clocks through the exchange of laser pulses. It combines ► *satellite laser ranging* with onboard measurements of the arrival time using a photo-detector on the satellite.

Latency

Delay in time after ► *differential GNSS* corrections are generated by the reference receiver or the network and the time they arrive and are applied by the user. Also: delay in the provision of precise orbit and clock products relative to the epoch of the GNSS observations used in their generation.

L-band

The band of the radio spectrum covering frequencies of 1–2 GHz.

Leap Second

An intentional time step of one second used to adjust ► *Universal Time Coordinated (UTC)* in order to ensure approximate agreement with UT1. An inserted second is called a positive leap second, and an omitted second is called a negative leap second.

Least-squares Ambiguity Decorrelation Adjustment (LAMBDA)

Method for the efficient computation of the integer least-squares ambiguity estimator. It makes use of a decorrelating ► *Z-transformation* to speed up the integer search and to compute an improved integer bootstrapping estimator as first approximation.

Least-squares Estimation

An estimation principle for solving an overdetermined system of equation by minimizing its (weighted) sum of squared residuals (the differences between input data and their estimated values).

Length-of-day

The rotational period of the solid Earth. The angular velocity of the solid Earth, and hence its period of

rotation, changes in response to the torques acting on it.

Light-time

The time that it takes an electromagnetic signal to pass the distance between the transmitter and receiver.

Linear Feedback Shift Registers (LFSR)

A shift register whose input binary state (0 or 1) is a linear function of its previous states. An n -stage shift register consists of n consecutive two-state stages (flip-flops) driven by a clock. At each pulse of the clock the state of each stage is shifted to the next stage in line to the right of the register. In order to convert the n -stage shift register into a sequence generator a feedback loop is incorporated, which calculates a new term for the left-most stage, based on the states of the n previous states. At the right-most stage of the register the generated sequence is output.

Line-of-sight (LOS)

Typically refers to the straight line between a GNSS satellite and the antenna of a GNSS receiver. The LOS signal is also referred to as the direct-path signal.

Line-of-sight Vector

Vector of unit length pointing from the receiver position to the satellite position.

Line Quality Factor

The ratio of the frequency of an atomic transition and the width of the resonance line. The \blacktriangleright *Allan Deviation* of an atomic clock is inversely proportional to the line quality factor, i. e., the clock stability improves with increasing quality factor.

Link budget

A budget of gains and losses in a telecommunication system. GNSS link budgets typically comprise the transmit power, the transmit and receive \blacktriangleright *antenna gains*, the \blacktriangleright *free-space loss* as well as atmospheric and cable losses.

Loading

The surface forces acting on the Earth from the movement of mass, and the deformation of the solid Earth in response to changing loading forces. Loading most commonly refers to water and atmospheric loading, but erosion and sedimentation can also produce measurable deformation where they are concentrated enough and involve a sufficiently large movement of mass.

Local Coordinates

Cartesian coordinates, often in a left-handed system, with the third axis along the local vertical.

Low Earth Orbit (LEO)

A satellite orbit with a representative altitude of about 300–1400 km that is commonly used for remote sensing missions.

Low-noise amplifier (LNA)

An electronic device to amplify electromagnetic signals, which is specifically designed to add only little noise and can thus be used for very weak signals.

M

Mapping Function

The ratio between the propagation delay through the atmosphere at a specified elevation angle and the propagation delay in the \blacktriangleright *zenith direction*.

Maser

A device producing coherent electromagnetic waves through *Microwave Amplification by Stimulated Emission of Radiation (MASER)*. Hydrogen Masers are used as atomic clocks, characterized by a very good short-term stability.

Master-auxiliary Concept (MAC)

Method of providing \blacktriangleright *Network RTK* corrections to the rover, where absolute corrections from one of the reference stations of the network (i. e., the master) and differential corrections from other (auxiliary) reference stations are transmitted to the rover.

Master Clock

A precision clock that provides timing signals to synchronize slave clocks as part of a clock network.

Master Station

A processing facility that collects measurements from multiple \blacktriangleright *reference stations* and makes determinations on the performance of the GNSS satellites (and perhaps the \blacktriangleright *ionosphere*). A master station may calculate orbits, satellite clock states, differential corrections, confidence bounds, and/or integrity evaluations.

Matched Filter

A technique to reveal a weak signal of known structure by \blacktriangleright *correlation* with a replica.

Maximum Likelihood Estimator (MLE)

A parameter estimator of which the outcome maximizes the respective likelihood function. A likelihood function is a parameter function that gives the probability or probability density for the occurrence of a corresponding observation or sample.

M-code

The modernized GPS military signals that are broadcast at center frequencies of 1575.42 MHz and 1227.6 MHz on Block IIR-M and subsequent GPS satellites.

Meaconing

A specialized spoofing attack in which an entire segment of radio frequency spectrum is captured and replayed. The term *meacon* is a portmanteau of *masking beacon*.

Mean Solar Time

An astronomical time scale that is based on the average length of the day, called the mean solar day. The length of an average day is different from a true or apparent solar day, due to daily variations, over the span of a year, in the Sun's apparent angular speed across the sky when viewed by an observer on Earth. Thus, the length of an average or mean solar day is used for a more uniform system of timekeeping.

Mean Tide System

A system (such as for coordinates) in which all

temporal tidal effects except the mean tidal effect have been removed.

Measurement-level Simulator

A simulator that generates pseudorange, carrier-phase, and/or Doppler measurements for computing the position and other relevant parameters in software. Typically used for testing the ability of a receiver or a third-party software to compute a correct solution.

Medium Altitude Earth Orbit (MEO)

An orbit with altitudes above ► *low Earth orbits* and below ► *geostationary orbits*. GNSS medium altitude Earth orbits are usually located at altitudes of about 20 000 km.

Meridian

An imaginary great circle on the celestial sphere that defines the plane that is parallel to the polar axis and contains a local vertical vector. The astronomic meridian plane contains the tangent to the local plumb line, and the geodetic meridian plane contains the local normal to the ellipsoid.

Minimal Detectable Bias (MDB)

The smallest bias that a test can detect for a given level of significance and ► *power*.

Minimum Descent Altitude or Height (MDA/H)

The lowest altitude or height to which a descent is authorized when an aircraft performs a ► *Non-precision Approach* procedure. Unlike a ► *Decision Altitude or Height* an aircraft must not descend below the MDA/H. The pilot may descend to this height and maintain it until reaching the missed approach point, where the missed approach must be initiated if the required visual references are not available.

Minimum Mean Penalty (MMP) Test

This ambiguity acceptance test penalizes certain ambiguity outcomes. The penalties (e.g., costs) are chosen by the user and can be made dependent on the application at hand. It is optimal in the sense that it minimizes the average of the assigned penalties.

Minimum Mean Square Error (MMSE) Estimator

The estimator that has the smallest mean-square-error of all estimators from a particular class. Often the class is restricted to the class of linear functions. The mean of the squared error is the mathematical expectation of the squared error.

Minimum Operating Performance Standards (MOPS)

A document describing the necessary requirements for an object to be awarded an airworthiness certificate. Federal advisory committees, operated by RTCA Inc., develop and document these requirements and the MOPS form the basis for Federal Aviation Administration (FAA) regulatory requirements.

Modified Julian Day Number

The ► *Julian day number* minus 2400000.5. The Modified Julian Date (MJD) has a starting point of midnight on November 17, 1858.

Modulation

The process of mapping ► *baseband* information to a high-frequency carrier for the purpose of transmission. This takes place by either varying

amplitude, phase, or frequency of the carrier signal to be transmitted.

Monitoring Station

► *reference station*

Moore's Law

An empirical relation named after a founder of Intel Co., who observed that the number of transistors in integrated circuits was almost doubled every 2 years.

Multi-function Satellite Augmentation System (MSAS)

A ► *satellite-based augmentation system (SBAS)* developed by the Japanese Civil Aviation Bureau to provide horizontal navigation throughout the Japanese airspace. It has provided safety-of-life service since 2007.

Multipath

The phenomenon whereby a transmitted GNSS signal is received at the receiver via multiple paths due to reflection and diffraction in addition to the direct-path signal. The received superposition of direct and non-direct-path signals (also known as the non-line-of-sight signals) results in errors in the signal tracking.

Multipath-to-direct (M/D) Ratio

The amplitude of a multipath signal relative to the direct-path signal.

Multiplexed Binary Offset Carrier (MBOC) Modulation

A variant of the ► *BOC* modulation, in which two sub-carriers of different frequency and different amplitude are used concurrently (composite BOC or CBOC) or in which the two sub-carriers have equal amplitude but alternate times slots (► *time multiplexed BOC or TMBOC*). The addition of the high-frequency sub-carrier component aims at an improved multipath resistance without a notable increase of the spectral width.

Multivariate Constrained-least-squares Ambiguity Decorrelation Adjustment (MC-LAMBDA)

A method for the computation of the integer least-squares ambiguity estimator subject to nonlinear constraints, based on the ► *Least-squares Ambiguity Decorrelation Adjustment (LAMBDA)* method.

N

Nadir

The direction towards the center of the Earth, i. e., opposite to the ► *zenith*.

Notice Advisory to Navstar Users (NANU)

Notifications issued by the ► *Global Positioning System (GPS)* to alert users of non-nominal situations such as signal outages or special operations that may result in service interruptions. Similar notifications (NAGUs, NAQUs) are provided by ► *Galileo* and the ► *Quasi-Zenith Satellite System*.

Narrow-lane Observable

A linear combination of carrier-phase observations on two frequencies that is formed as the sum of the individual phase values expressed in cycles. It exhibits

a small effective wavelength, e.g., 10.7 cm for GPS L1 and L2.

Navigation

The estimation process for determining the system ► *pose* as it maneuvers. Also, the process of determining and implementing a trajectory for an autonomous system.

Navigation Message

The set of auxiliary parameters such as satellite orbit and clock information, ionospheric correction data, and time offset parameters, which are transmitted by a GNSS satellite to enable computation of a position fix from the ► *pseudorange* measurements.

Navigation Message Authentication

A GNSS signal authentication technique whereby unpredictable (to the public) but verifiable data are inserted into a GNSS signal's navigation data stream. The properties of unpredictability and verifiability can be achieved by means of a digital signature, which is generated by a secret (private) key but can be verified by a public key. Besides injecting verifiable randomness into the navigation data stream, the digital signature serves to authenticate all data in the stream as originating with the holder of the private key (e.g., the control segment for a particular GNSS constellation).

Navigation System Error (NSE)

The difference between the true position and the estimated position. This error is linked to the navigation system providing the position estimation.

NeQuick

An empirical model used to describe the ionospheric electron density. A specific version known as NeQuick-G has been adopted by the Galileo system. Here, the modeled electron density is integrated along the signal path and used to correct single-frequency code measurements. The NeQuick-G model parameters are broadcast by the Galileo satellites with at least a daily update interval.

Networked Transport of RTCM via Internet Protocol (NTRIP)

Protocol for the streaming of ► *differential GNSS* corrections in ► *RTCM* format via the Internet.

Network RTK (NRTK)

A ► *differential GNSS* positioning technique that extends the operational distance of the ► *real-time kinematic (RTK) positioning* technique from 10 km to about 100 km from the reference receiver. This is realized by employing a network of reference receivers that produces precise correction models of the distance-dependent biases (atmosphere, orbit).

Neuman-Hofman (NH) Code

A specific type of ► *secondary code* used in ► *pilot signals* of the modern GNSS navigation signals.

Noise

Random fluctuations in a measured signal.

Non-directional Beacon (NDB)

A radio beacon operating in the medium or low-frequency bandwidths used for aircraft navigation. NDBs transmit a signal of equal strength

in all directions. ► *Automatic Direction Finding (ADF)* equipment on board aircraft uses bearings from NDBs for navigation purposes.

No-net-rotation

The conventional requirement that subsequent realizations of a reference system are parallel.

Non-precision Approach (NPA)

An aircraft approach procedure using lateral guidance to bring the aircraft to a point where the runway is in view and a visual landing can be performed. NPA procedures do not include vertical guidance and include a ► *Minimum Descent Altitude/Height*.

Notice to Airmen (NOTAM)

A notice containing information concerning the establishment, condition, or change in any aeronautical facility, service, procedure, or hazard, the timely knowledge of which is essential to personnel concerned with flight operations.

Null Hypothesis

► *Hypothesis testing*

Numerically Controlled Oscillator (NCO)

A digital signal generator used to generate a harmonic wave of desired frequency, and, optionally, phase. It is used inside a ► *phase lock loop* or ► *frequency lock loop* to track the incoming carrier signal and to measure its phase and ► *Doppler shift*.

Nutation

An oscillation of the Earth's axis about its mean position, which is superimposed on precession and driven by the luni-solar gravitational torque. The nutation movement is the resultant of oscillations with different periods of which the more important is 18.6 years, corresponding to one rotation of the Moon's ascending node.

O

Obliquity

The angle between the equatorial and orbital planes of some body. For the Earth, the angle between the plane of the Earth's equator and the plane of the Earth's orbit about the Sun. The mean obliquity of the ecliptic is about 23.4°.

Obliquity Factor

The ratio between a slant path passing through a thin slab above the Earth to a vertical path passing through the same location in the slab. The obliquity factor (or ► *mapping function*) is used to convert a vertical ionospheric delay estimate into a slant delay estimate corresponding to the elevation angle of the path. For the thin shell model that assumes an ► *ionosphere* 350 km above the Earth, the obliquity factor ranges from 1 (for a satellite directly over head) to just over 3 (for a satellite close to the horizon).

Observation Session

The length of time a non-moving receiver makes sufficient observations (adequate change in satellite–receiver geometry) for reliable ► *static positioning*. It may be many hours in length for a long ► *baseline* and/or an ► *ambiguity float solution*, to as

short as a few seconds if the ambiguities have already been resolved.

Occultation

► *Radio Occultation*

Orbit

The trajectory of a natural or artificial satellite around a central body – the Sun in the case of planets in the solar system, or the Earth in the case of artificial Earth-orbiting satellites.

Orbital Plane

The plane to which the motion of a satellite is confined in a central gravity field. In the presence of perturbations an ► *osculating* ► *orbit plane* is defined by the instantaneous position and velocity vector.

Orbit Determination

The process of estimating a set of parameters (including the initial position and velocity as well as various force model parameters) describing the future motion of a satellite through a corresponding dynamical trajectory model.

Orbit Perturbations

Deviations of the orbital motion of a satellite around a central body from an idealized ► *Keplerian orbit*. Such perturbations are, e.g., caused by the aspherical gravity field of the Earth, the luni-solar gravitational attraction, atmospheric drag, and solar ► *radiation pressure*.

Oscillator

A device providing a periodical signal with a given frequency.

Outlier

A measurement that, relative to the assumed measurement noise probability distribution, is so rare that its validity is questionable.

Overbound

A probability distribution whose likelihood of having an error magnitude greater than some value is at least as large as the likelihood that the true error magnitude is greater than that value. The overbound is most commonly specified as a 1-sigma value for a zero-mean Gaussian distribution. The overbound is used to conservatively describe a true error distribution.

Overall Model Test

A test that tests the null hypothesis (► *hypothesis testing*) against the most relaxed alternative. This test is used for detecting unspecified modeling errors in the null hypothesis.

Overlay Code

► *Secondary code*

P

Parametry Zemli 1990 (PZ-90)

An Earth model maintained by the Military Topography Agency of the General Staff of the Armed Forces of the Russian Federation. The definition of PZ-90 comprises fundamental geodetic constants, parameters of the Earth's ellipsoid, and the Earth's

gravity field parameters, as well as the geocentric reference system, which is defined in accordance with common conventions of the International Earth Rotation and Reference Systems Service (IERS) and Bureau International de l'Heure (BIH). PZ-90 serves as the ► *datum* for ► *GLONASS* ► *single-point positioning*.

Partial Ambiguity Resolution (PAR)

Ambiguity resolution when only a part of the ambiguities are resolved as integers. PAR can be applied in case the GNSS model is not strong enough to enable successful ambiguity resolution of all ambiguities. PAR is usually applied after the ambiguities have been re-parameterized with a proper ► *Z-transformation*.

Parts-per-million (ppm)

A relative accuracy measure defined as the ratio of accuracy to ► *baseline* length scaled by one million. For example, 1 ppm is 1 cm accuracy between two GNSS receivers separated by 10 km, or 10 cm over 100 km, etc. Parts-per-billion (ppb) is obtained as ppm × 1000. For example, 10 ppb is 1 cm accuracy between two GNSS receivers separated by 1000 km.

P-code

The precision (P) ranging code modulated onto both L1 and L2 carriers in GPS and GLONASS. The GPS P-code is referred to as the Y-code if it is encrypted. In the case of GPS, only authorized receivers are capable of directly tracking the Y-code. Civilian receivers use ► *semi-codeless tracking* techniques to obtain P(Y)-code pseudorange measurements.

Performance-based Navigation (PBN)

► *Area navigation* based on performance requirements for aircraft operating along an air traffic service route, on an instrument approach procedure or in a designated airspace. Airborne performance requirements are expressed in navigation specifications in terms of ► *accuracy*, ► *integrity*, ► *continuity*, and functionality needed for the proposed operation in the context of a particular airspace concept. Within the airspace concept, the availability of GNSS signal-in-space (SIS) or that of some other applicable navigation infrastructure has to be considered in order to enable the navigation application.

Perifocal Coordinates

Position of a celestial body relative to its orbital plane and the line of apsides.

Perigee

The closest point of an artificial satellite's orbit around the Earth.

Phase Bias

The signal hardware delay at receiver and transmitter side associated with the carrier phase signal generation.

Phase Center Offset (PCO)

The separation vector between the ► *antenna reference point* and the ► *antenna phase center*.

Phase Center Variation

Deviations of the antenna radiation pattern from a perfect sphere about the ► *antenna phase center*.

Phased-array Antenna

An array of antennas in which the relative phases of the signals coming to the antennas are combined in such a way that the effective radiation pattern of the array is reinforced in a desired direction and suppressed in undesired directions.

Phase-range Corrections

Corrections determined at a (network of) reference receiver(s) that are transmitted to a rover receiver in order to enable carrier-phase-based ► *differential GNSS* or ► *real-time kinematic positioning*.

Phase Lock Loop (PLL)

A controller used to align the phase of a carrier replica inside a GNSS receiver with that of the incoming signal. It comprises a ► *numerically controlled oscillator*, a phase ► *discriminator* that senses the instantaneous tracking error, and a loop filter that provides a smoothed estimate of the phase error for feedback to the NCO.

Phase Unlock

Failure to maintain phase tracking lock in a carrier phase tracking loop, resulting in a single ► *cycle slip* or a succession of cycle slips.

Phase Velocity

The speed of propagation of the carrier signal of an electromagnetic wave at a single frequency. It describes the velocity of movement of the phase front.

Phase Wind-up

The change in the received signal phase due to rotational changes in the relative orientation between receiver and transmitter antenna in the direction of signal propagation.

Pilot Signal

A GNSS signal component that does not contain data and is only modulated with a ranging code. Pilot signals allow for extended integration times for improved tracking sensitivity and robustness. Often, a ► *tiered code* is used for pilot signals in which a medium length primary ranging code is combined with a short ► *secondary code*.

Pivot Receiver/Satellite

Receiver or satellite that is selected as reference for forming ► *between-receiver* or ► *between-satellite differences*, respectively.

Plate Boundary Zone

The boundaries between tectonic plates are observed to be diffuse and often involve deformation spread out over regions hundreds to even 1000 km wide. Within a plate boundary zone, several active faults may take up the relative plate motion between the major plates, sometimes with large undeforming regions inside the plate boundary zone.

Plate Motion

The surface of the Earth is broken up into a set of tectonic plates, which are rigid except near their edges and which are all moving relative to each other. Plate

motions are steady with time, changing only over long timescales, like hundreds of thousands of years. Plate motions are described in terms of rotations on the surface of a sphere about a geocentric axis.

Polarization

The direction of the oscillation of the electromagnetic field as a function of time. GNSS signals are circularly polarized, i. e., the field vector rotates along the propagation direction.

Polar Motion

The motion of the ► *celestial intermediate pole* with respect to the crust and mantle of the Earth. Can also refer to the motion of some other pole, such as the rotation pole, with respect to the crust and mantle of the Earth.

Pose

The position and attitude of an entity.

Positioning

Determination of the position coordinates of a location (in a reference frame) by means of measurement techniques in which the instrument is either placed on the position to be determined, or where the instrument measures the location of which the position has to be determined.

Position Dilution of Precision (PDOP)

► *Dilution of Precision*

Postseismic

Immediately after an earthquake. Transient deformation processes are observed to occur immediately after large earthquakes. These processes may cause very rapid deformation for a short time after the earthquake, and the rate of deformation decays with time back to a background, interseismic rate.

Power (statistical)

One minus the ► *probability of missed detection*.

Preamble

A well-defined bit sequence used to identify the start of data frames in the GNSS navigation message.

Precession

The slow variations in the directions of the Earth's instantaneous spin axis and of the ► *vernal equinox* relative to the celestial sphere due to the gravitational actions of the Sun, Moon, and planets on the Earth's orbit and its non-spherical shape and non-homogeneous constitution.

Precise Orbit Determination (POD)

A technique that combines methods and strategies to derive precise (typically of sub-decimeter accuracy) satellite positions using either a dynamic (relying on accurate modeling of forces acting on a satellite) or kinematic (trajectory through epoch-wide representation) approach.

Precise Point Positioning (PPP)

A technique that makes use of pseudorange and carrier phase observations of only a single receiver along with precise orbit and clock information of the GNSS satellites for determining the position of the receiver antenna.

Precise Point Positioning Real-time Kinematic (PPP-RTK)

Extension of the ► *precise point positioning (PPP)* technique by including satellite phase bias corrections such that the single-station carrier-phase ambiguities can be resolved to integers and, consequently, the PPP precision can be improved to centimeter level.

Precise Positioning Service (PPS)

One of two services provided by GPS that is intended for authorized (e.g., military) users only and based upon the ► *P(Y)-code* signals on two frequencies, GPS L1 and L2.

Precision

A measure for the reproducibility of measured or estimated quantity when measurement or estimation is repeated under similar circumstances.

Precision Approach

an instrument approach procedure using precise lateral and vertical guidance flown to a ► *decision altitude/height*.

Prediction (statistical)

Estimation of the outcome or realization of a random variable or vector. An observable random vector is used to guess the outcome of another non-observed random vector. The non-observable vector may comprise model parameters to be predicted in time or in space, but also signal and/or noise parameters.

Probability of False Alarm

The chance of rejecting the null hypothesis (► *hypothesis testing*) while it is true. It is also known as the significance level and it is usually denoted by α .

Probability of Hazardous Missed Detection

The product of the ► *probability of hazardous occurrence* and ► *probability of missed detection*

Probability of Hazardous Occurrence

Probability of having the outcome of the GNSS parameter estimator lie outside a pre-defined, non-hazardous parameter region. This probability increases as the influential ► *bias-to-noise ratio* gets larger.

Probability of Missed Detection

The chance of not rejecting the null hypothesis (► *hypothesis testing*) while it is false. It is usually denoted by β . This probability gets smaller as the testable ► *bias-to-noise ratio* gets larger.

Proper Time

The time scale associated with an observer at rest in a local frame.

Protection Level

The maximum possible positioning error that may be present for a navigation system at the current time, for a specified probability level. Usually, the protection level is compared to a corresponding ► *alert limit* to determine whether the navigation system meets the operational requirements at that time.

PRN Number

A number used to identify a GPS satellite based on the transmitted signal. More specifically, the PRN number denotes the serial number assigned to the ► *C/A-code* ► *pseudo-random noise sequence*.

Pseudolite

A device that transmits GNSS-like ranging signals from a known location to augment or replace the signals broadcast by GNSS satellites. The word is a contracted form of the composite term *pseudo-satellite*.

Pseudo-random (PR) Binary Sequence

► *Pseudo-random noise*

Pseudo-random Noise (PRN)

A quasi-random bit sequence of limited length with good cross and autocorrelation properties. PRN sequences are commonly used as ranging codes in GNSS systems.

Pseudorange

A distance-like measurement obtained from the time difference between transmission and reception of a radio signal and the known speed-of-light. Due to time offsets between the local clocks measuring the two times, the measurement differs from the true distance and includes a contribution related to these clock offsets. It is hence called a *pseudorange*.

Pseudorange Corrections

Corrections determined at a (network of) reference receiver(s) that are transmitted to a rover receiver in order to enable code-based ► *differential GNSS (DGNSS)* positioning.

Pull-in Region

Region in which every float ambiguity vector is pulled to the same integer vector. Pull-in regions are translational invariant regions that cover the ambiguity space without gaps and overlap. The shape of the pull-in region is defined by the type of integer estimator chosen.

P-value

A measure of strength-of-evidence on which the ► *hypothesis testing* decision to *reject* or *not reject* is made. Given the data, it is the smallest significance level at which the test rejects the null hypothesis.

P(Y)-code

A 10.23 MHz chipping rate, spread-spectrum signal broadcast by the GPS satellites on two frequencies, 1575.42 MHz and 1227.6 MHz. The precision (P)-code is unencrypted. For many years, the P-code has been encrypted into what is referred to as the ► *Y-code*. In common usage, P(Y)-code refers to the 10.23 MHz chipping rate GPS signals whether they are being broadcast encrypted or unencrypted.

Q**Quadrature (Q) Component**

A signal component transmitted with a 90° phase shift relative to the ► *in-phase component* of a compound navigation signal.

Quadrature Phase Shift Keying (QPSK)

A modulation scheme for radio navigation signals, in which two superimposed carriers with a 90° shift (known as in-phase and quadrature channel) are each modulated with a binary signal using ► *binary phase shift keying (BPSK)*.

Quadrifilar Helix

A GNSS antenna made up of four wires arranged in a fractional-turn helix configuration and fed with progressive quadrature phase shifts.

Quantization

The process of converting a signal defined on a continuous range of values to one on a finite range of discrete values. The analog radio frequency signal received by a GNSS signal may be quantized to two, three, four, or more discrete quantization levels. Higher quantization resolution resulting from a larger number of quantization levels reduces signal distortion due to quantization.

Quartz Crystal Oscillator

A harmonic signal generator comprised of an specially cut quartz crystal device in a tuned circuit designed to produce a specific frequency signal. Design variations are employed to compensate for environmental effects on the crystal device that may cause frequency changes.

Quasi-Zenith Satellite System (QZSS)

A regional Japanese navigation system, which uses slightly eccentric ► *inclined geosynchronous orbits* to ensure that at least one satellite is always visible at high elevations for users in the service area.

Quaternion

A real-valued, four-component entity that extends the space of complex numbers by defining a hypercomplex mathematical object. A subset of the space of quaternions (quaternion with unit norm) can be used to parameterize a rotation.

R**Radiation Pressure**

The pressure caused by the absorption or reflection of photons impinging on the surface of a satellite. For GNSS satellites with large solar panels, radiation pressure is a dominant source of ► *orbital perturbations*. Aside from the direct solar radiation pressure, reflected solar radiation of the Earth (► *albedo*) or the Earth's infrared radiation contribute to the total acceleration.

Radio-determination Satellite Service (RDSS)

A service defined by the International Telecommunications Union (ITU) for location determination and reporting of mobile users using radio signals in the L-band (uplink) and S-band (downlink).

Radio Occultation

A radio technique that measures the change of radio wave parameters such as signal strength and phase at grazing incidence when the radio wave continuously approaches the surface of a planet until the radio wave finally disappears. The technique has been widely used to explore planetary atmospheres such as Venus and Mars. The GNSS radio occultation technique uses the changing refraction of GNSS signals while approaching the Earth's atmosphere to

retrieve vertical profiles of the electron and neutral gas densities of the ionosphere and troposphere, respectively.

Radome

A dome that covers, e.g., a ► *choke-ring antenna*, as typically used for geodetic surveying applications or at a ► *reference station*.

Ranging Code

A binary sequence modulated on a carrier wave to enable ► *(pseudo)range* measurements. GNSS uses ► *pseudo-random noise (PRN)* sequences as ranging code.

Ratio Test

An ambiguity test to decide whether or not to accept the estimated ► *integer ambiguity vector*. The test is based on the ratio of two quadratic forms, measuring the closeness of the float ambiguity vector to the estimated integer vector and the next nearest integer vector.

Ray Tracing

The reconstruction of the signal path through different media.

Real-time Kinematic (RTK) Positioning

► *Differential GNSS* positioning technique that is driven by carrier-phase data based on a baseline set up between a reference and rover receiver. Essential to high-precision RTK positioning is carrier-phase integer ambiguity resolution. Code (pseudorange) data are used in addition to the phase data to strengthen the RTK positioning model. For sufficiently short baselines (e.g., less than 10 km) the differential atmospheric biases can be neglected and very fast integer ambiguity resolution is feasible.

Rebroadcast Test

Testing a receiver by broadcasting GNSS signals (usually simulated) toward the device under test from one or more transmitter positions. Typically conducted when the antenna is integrated into the device under test, or when it is necessary to include the antenna in the test chain.

Receiver Autonomous Integrity Monitoring (RAIM)

A testing procedure whereby the redundant observations available at the GNSS receiver are autonomously processed to monitor the integrity of the GNSS signals with the purpose of providing relevant warnings.

Receiver Independent Exchange Format (RINEX)

The receiver independent exchange format is an ASCII-based format for GNSS observation and navigation data, as well as meteorological data.

Record-and-playback System

A system capable of recording received GNSS signals as intermediate frequency samples and, at a later time, up-converting the replayed signals for input to a GNSS receiver.

Redundancy

The total number of available observations minus the number of observations that are strictly needed to solve the system of equations. For a linear system of

observation equations it equals the difference between the number of observations and the rank of the system matrix.

Reference Ellipsoid

That ► *ellipsoid* adopted for a particular national or global ► *datum* or ► *reference frame*, such as the ITRF. The ► *Geodetic Reference System 1980 (GRS80)* ellipsoid is an internationally recognized reference ellipsoid.

Reference Frame

The realization of a ► *reference system* by means of coordinates of ► *control points* or ground marks that are accessible directly by occupation or by observation; for example, the ► *International Terrestrial Reference Frame (ITRF)*.

Reference Station

A GNSS receiver at a precisely surveyed antenna location (i. e., with known coordinates expressed in a ► *reference frame*), whose measurements are used to monitor, and possibly correct, any observable satellite signal errors. The reference station can act as the coordinate fixed point for baseline solutions or relative positioning determination with GNSS techniques such as ► *differential positioning (DGNS)* or ► *real-time kinematic (RTK)* positioning.

Reference System

A set of prescriptions and conventions together with the modeling required to define at any time a triad of Cartesian coordinate axes.

Reflectometry

Method to establish properties of a reflecting surface by comparing the properties of incident (or a replica) and reflected electromagnetic signals.

Refraction

The deflection of GNSS signals in the Earth's atmosphere.

Regional Argumentation

The provision of additional parameters derived from regional GNSS observations to support ► *Precise Point Positioning (PPP)*

Relative Positioning

► *Differential GNSS*

Required Navigation Performance (RNP)

A form of ► *Area Navigation (RNAV)* with the addition of an on-board performance monitoring and alerting capability.

RF-level Simulation

A simulator that generates radio-frequency (RF) signals similar to those expected at the input of an antenna. Typically used for conductive testing, but rebroadcast testing is also possible.

Right Ascension

The longitude in a ► *celestial coordinate system*. Right ascension is the angle between the reference direction (at or close to the ► *vernal equinox*) and the projection of a bodies position on the equatorial plane, measured in an eastern direction.

RINEX

► *Receiver independent exchange format*

Rodrigues Vector

A vector of ► *attitude parameters* derived from the elements of a ► *quaternion*.

Rotation Poles

The two points, the north rotation pole and the south rotation pole, defined by the intersection of the Earth's rotation axis with the surface of the Earth.

RTCM Message Format

Standardized format for the exchange of GNSS observations, ephemerides, and correction data as defined and published by the Radio Technical Commission for Maritime Services Special Committee 104 (RTCM SC-104).

Rubidium Atomic Frequency Standard

A signal generator that produces a stable signal based on optically pumping the hyperfine frequency of Rubidium 87 at 6.834682611 GHz, where the Rubidium is suspended in a gas cell.

S

S-band

A part of the spectrum of electromagnetic waves with carrier frequencies in the range of 2–4 GHz.

Sagnac Correction

In the context of GNSS, the Sagnac or Earth-rotation correction denotes a (non-relativistic) correction of the satellite positions that must be applied in the computation of the navigation solution when working in a rotating, Earth-fixed reference frame, to properly account for the Earth's rotation during the signal propagation time.

Sample Rate

The frequency at which a signal is measured.

Satellite-based Augmentation System (SBAS)

A wide area differential GNSS augmentation system using a regional monitoring network to collect data from core constellations and providing a navigation message to users via satellites in ► *geostationary orbit*. Examples include the US ► *Wide Area Augmentation System (WAAS)*, the ► *European Geostationary Navigation Overlay Service (EGNOS)*, the Japanese ► *Multi-function Satellite Augmentation System (MSAS)*, and the Indian ► *GPS Aided GEO Augmented Navigation (GAGAN)* system.

Satellite Laser Ranging (SLR)

A geodetic technique that provides distance measurements between satellites and a ground station based on the signal turn-around time of laser pulses.

Scintillation

Temporal fluctuations in phase and intensity caused by electron density irregularities along a transionospheric signal's propagation path. Scintillation effects may lead to severe signal fading (e.g., deep power fades > 15 dB) associated with loss of lock and extremely enhanced phase noise.

Search and Rescue (SAR)

A secondary mission for some GNSS constellations, which involves the detection of internationally standardized distress signals from emergency beacons

and relaying of this information to government authorities.

Second

The duration of 9 192 631 770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the cesium-133 atom. The definition was added to the International System (SI) of units in 1967.

Secondary Code

A short binary pattern that is applied to subsequent repetitions of a fast, medium-length, primary spreading code to form a long ► *tiered code* that enables long integration times. Also referred to as an overlay or synchronization code.

Selective Availability (SA)

An intentional degradation of the clock phase to limit the accuracy of the ► *standard positioning service* of the ► *Global Positioning System (GPS)* for civil users to approximately 150 m. Selective availability was finally abandoned by presidential order in May 2000.

Semi-codeless Tracking

A special technique to track the encrypted ► *Y-code* signal of the GPS satellites without full knowledge of the signal. It is based on the assumption that the Y-code results from the known ► *P-code* by multiplication with an unknown low rate (≈ 500 kHz) *W-code*.

Semi-kinematic Positioning

► *Differential GNSS* positioning technique in which carrier-phase (and pseudorange) data are collected for a rover receiver that is moving with respect to a stationary reference receiver. The rover receiver collects data during a short time (a few minutes) and then moves to the next point, continuously tracking the signals. To avoid a long observation time span during which the rover cannot move (as with conventional static positioning), special measurement procedures have been developed (i. e., revisiting of stations, starting from a known baseline, with antenna swap).

Semi-major Axis

Half the *large diameter* of an ellipse, i. e., the radius of an encompassing circle. For an elliptic satellite orbit, the semi-major axis denotes the mean value of the minimum and maximum orbital distance from the central body.

Shapiro Effect

The gravitational time delay experienced by an electromagnetic signal due to the presence of a massive body close to the signal transmission path.

Shielding Chamber

A chamber or cabinet designed to contain radio-frequency signals. Typically used for broadcast testing of radio-frequency equipment when the device under test must be shielded from external signals and/or when the broadcast signals are in a protected or restricted band.

Sidereal Day

The interval of time between two consecutive upper transits of the ► *vernal equinox* across some

► *meridian*. The mean sidereal day is 86 164.09054 s long and is a measure of the rotation of the Earth with respect to the stars.

Sidereal Time

The time associated with Earth's rotation relative to the celestial sphere, where 15° of rotation equals 1 h of sidereal time.

Signal-in-space Range Error (SISRE)

The user range error contributed by both the ► *space segment* and ► *ground segment*, but excluding ionosphere, troposphere, multipath errors, and receiver noise contributions. Usually applied to define the navigation service quality of the system itself.

Signal-to-noise Ratio (SNR)

A signal power to noise power ratio. It compares the level of a desired signal to the level of background noise.

Signal-to-interference-plus-noise ratio (SINR)

A signal power to noise-plus-interference power ratio.

Significance Level

The ► *probability of false alarm*.

SINEX

► *Solution independent exchange format*

Single-difference

► *Between-receiver difference* or ► *between-satellite difference* of GNSS observations and parameters.

Single Point Positioning (SPP)

An absolute GNSS positioning technique that is based on pseudorange measurements of at least four satellites with known positions and clocks offsets.

Slant Total Delay

The extra time needed for a signal propagating through the neutral atmosphere in a given (slant) direction compared to the propagation time in vacuum. It is often expressed in units of length, using the speed of light in vacuum for the conversion. For practical reasons, the slant total delay (STD) is divided into a slant hydrostatic delay (SHD) and a slant wet delay (SWD). A special case is the delay in the zenith direction. This zenith total delay (ZTD) is also divided into a zenith hydrostatic delay (ZHD) and a zenith wet delay (ZWD). The ZHD is approximately 2.3 m at sea level and proportional to the ground pressure. The ZWD can be anything between 0–40 cm, depending on the climate zone and the specific weather conditions.

Software Defined Radio (SDR)

A radio communication system implemented by means of software running on a processing system instead of typical implementation in hardware.

Solar Day

The interval of time between two consecutive transits of the Sun across some ► *meridian*. The nominal solar day is 86 400 s long and is a measure of the rotation of the Earth with respect to the Sun. The length of the solar day differs from that of the ► *sidereal day* by about 4 min because of the orbital motion of the Earth about the Sun.

Solar Radiation Pressure

The non-gravitational force acting on a satellite due to the direct radiation of the Sun (► *radiation pressure*).

Solar Radio Burst

An intense outburst of radio emissions from the Sun, with spectral power ranging from 3 MHz to above the L-band. A burst is typically associated with solar flares, which are caused by the acceleration of electrons in the solar atmosphere and whose rate of occurrence follows the 11-year sunspot cycle.

Solution Independent Exchange Format (SINEX)

An ASCII-based format for normal equation or variance/covariance matrices and related information. SINEX files computed by different analysis/combination centers are, e.g., the input for the computation of the ► *International Terrestrial Reference Frame*.

Space Segment

A key part of a ► *global navigation satellite system*, comprising the constellation of satellites with proper orbital geometry which transmits the navigation signals.

Space Vehicle Number (SVN)

A consecutive number assigned to different satellites of the Global Positioning System (GPS). Other than the pseudo-random noise (PRN) number the SVN is unique for a given spacecraft and does not change throughout its lifetime.

Space Weather

characterizes the energy, intensity, and composition of the electromagnetic and corpuscular solar radiation, galactic cosmic rays, and the associated state and coupling processes of the magnetosphere, ionosphere/plasmasphere, and thermosphere.

Specific Force

The non-gravitational force per unit of mass.

Spoofing

The act of generating a signal whose structure adheres closely enough to a GNSS signal specification that it can be misconstrued by a GNSS receiver as authentically broadcast by a GNSS satellite. Spoofing can be intentional, as in a deliberate attempt to manipulate the position, velocity, or time readout of a target GNSS receiver, or unintentional, as in an errant GNSS simulator or repeater signal that could be misinterpreted as originating from a GNSS satellite.

Standard Positioning Service (SPS)

One of two services provided by GPS that is intended for civilian use and is based upon the ► *C/A-code* signal on one frequency, GPS L1 (1575.42 MHz).

Static Positioning

Estimation of a single set of coordinates for a non-moving receiver from observations covering an extended observation data collection session (typically 1 or more hours). Referred to as rapid static positioning when the ► *observation session* is of the order of a few tens of minutes.

Stochastic Orbit Parameter

Empirical parameters such as accelerations or impulsive velocity increments that are introduced into

the equation of motion of a satellite and adjusted in the orbit determination process to compensate for imperfections of the employed force model.

Stratosphere

is the layer in the Earth's atmosphere above the ► *troposphere*. It starts in about 8–13 km, but the actual value depends on the weather conditions and varies systematically with the latitude and the season. The top is at a height around 50 km.

Surface Acoustic Wave (SAW) Filter

A bandpass filter for radio frequency signals based on an electromechanical device that converts electrical signals to a mechanical wave and then back to electrical signals.

Surplus Satellite

An extra satellite in a GNSS constellation that is not operational (e.g., because it is older than its design life and suffers some performance degradation), but could be reactivated if needed.

System of Differential Correction and Monitoring (SDCM)

A ► *satellite-based augmentation system (SBAS)* being developed by the Russian Federation to provide horizontal and vertical navigation throughout Russia.

T

Terrestrial Time (TT)

Terrestrial Time is the relativistic timescale that replaced ► *Ephemeris Time (ET)* as the time reference for apparent geocentric ephemerides (► *Dynamical Time*). For practical purposes both time scales may be considered to be equivalent. Its origin is defined by the following relation to TAI: $TT = TAI + 32.184$ s on January 1, 1977, 0 h TAI. TT is a theoretical ideal, which real clocks can only approximate; its best realization is TT(BIPM) provided on a yearly basis by the BIPM from a set of atomic clocks also used for TAI.

Testable Bias

Bias that propagates into the test statistic; such bias lies in the orthogonal complement of the design matrix range. A non-influential bias is always testable.

Test statistic

A function of the observables that is used to test hypotheses.

Thermal Noise

Broadband noise originating in an electrical conductor due to the random thermal motion of electrons. In a radio system such as a GNSS receiver, thermal noise originates primarily in the first amplifiers through which received signals pass. It can be accurately modeled as spectrally flat with an intensity proportional to temperature and having a Gaussian amplitude distribution.

Tides

Earth deformations induced by the luni-solar gravitational attraction and resulting in periodic ground motions (body tides) of several tens of centimeters, inducing periodic displacement of the liquids at the surface of the Earth (ocean tides).

Besides changing the position of points on the surface of the Earth, tides also cause small variations of the Earth's gravity field.

Tide-free System

A system (such as for coordinates) in which all tidal effects have been removed.

Tiered Code

A combination of a primary ranging code and a short ► *secondary code* commonly used in ► *pilot channels* of modern GNSS signals.

Time Division Multiple Access (TDMA)

A multiple access scheme, where channel users (satellites) occupy the complete available bandwidth but at different times, i. e., transmitting in turn in assigned time slots.

Time Multiplexed Binary Offset Carrier (TMBOC) Modulation

A modulation in which different ► *binary offset carrier modulations* pulse shapes are used for different chips of the pseudo-random binary sequence. For example, a mixture of BOC(1,1) and BOC(6,1) modulations is used for the ► *pilot component* of the GPS L1 civil (L1C) signal.

Time Scale

A continuous realization of a (conventional) reference frequency

Time-to-first-fix (TTFF)

The time between activation of a GNSS receiver and the first computation of a navigation solution. It is determined by the time required to search and for a sufficient number of satellites, to reliably track them and to decode the relevant parts of the navigation message. TTFF may vary from a few seconds for a ► *hot start*, to tens of seconds for a ► *warm start*, or even up to a few minutes for a receiver ► *cold start*.

Time-to-alert (TTA)

A maximum time allowed when a system that was previously declared safe for use can no longer assure that it meets all its integrity requirements for a given operation.

Time Transfer

The transfer of a precise reference time needed for remote synchronization. In scientific metrology, the time transfer is also used for remote comparisons of atomic clocks.

Timing Group Delay (TGD)

Scaled value of a satellite ► *differential code bias* as transmitted in the satellite's navigation message.

Tomography

Tomography refers to imaging of an object by penetrating waves whose modifications are measured after leaving the target. As an example, ground and space-based GNSS signals can be used to image the electron density distribution in the ionosphere and the water vapor distribution in the troposphere by measuring code and/or carrier phase changes.

Total Electron Content (TEC)

Integral of the electron density along a given ray path through the ionized atmosphere. Since each ray path has a specific geometry in concrete applications, TEC

must be specified by both ends of the ray path and the elevation angle. In ground-based GNSS applications it is convenient for distinguishing between the slant TEC (STEC) along the entire ray path and the geometry free vertical TEC (VTEC) that describes the vertical electron content from the bottom of the ionosphere up to GNSS orbit heights and is commonly used as reference. TEC is usually measured in units of 10^{16} electrons per m^2 that is equivalent to 1 TEC unit (TECU).

Total Station

A survey instrument set up on a tripod over a ground mark that electronically measures the horizontal and vertical angles of the telescope when pointed at a target, as well as the distance to a reflecting prism (or reflective surface) using an infrared laser. Used to transfer geodetic coordinates from the ground mark to a target.

Tracking

The continuous alignment of a replica signal generated inside a GNSS receiver to the received signal. Based on the tracking process, which is initiated after the initial signal ► *acquisition*, measurements of code delay, carrier phase, and carrier can be obtained by the receiver.

Tracking Loop

A controller used to align a replica of the carrier or ranging code in a GNSS receiver with the incoming signal. A ► *phase lock loop (PLL)* or ► *frequency lock loop (FLL)* for carrier tracking is combined with a ► *delay lock loop (DLL)* to track the ranging signal.

Traveling Ionospheric Disturbance

Ionospheric perturbation of electron density characterized by a horizontal scale length of a few hundred kilometers that travel at velocities in the order of a few hundred meters per second. TIDs are often generated by ► *space weather* events, in particular at high latitudes due to the interaction of solar wind with the Earth's magnetosphere, causing enhanced ionization and heating. Here the enhanced solar energy input causes perturbation-related thermospheric winds and electromagnetic forces from the magnetosphere, which may move plasma perturbations, e.g., towards lower latitudes.

Traveling Wave Tube Amplifier (TWTA)

A traveling wave tube integrated with a regulated power supply and protection circuits used to produce high-power radio frequency signals.

Triple-difference

The time difference of ► *double-difference* GNSS observations.

Troposphere

is the lowest layer of the Earth's atmosphere, where the temperature on the average decreases with height. It ends at the tropopause, which is located in the range from 8 km to 13 km, depending on the weather conditions, and varies systematically with latitude and season. The troposphere contains the weather, e.g., clouds and precipitation.

Tropospheric Refraction

Describes the signal propagation delay and bending induced by the electromagnetic neutral part of the atmosphere (► *slant total delay*). The ► *wet delay* and ► *hydrostatic (dry) delay* components are typically separately modeled or accounted for in GNSS measurement processing.

Two-body Problem

The task of calculating the motion of two bodies under the influence of their mutual gravitational attraction. The two-body problem is a simplified representation of the motion of a satellite around the Earth, where all perturbations are neglected. The solution of the two-body problem is also termed a ► *Keplerian orbit*.

Two-way Satellite Time and Frequency Transfer (TWSTFT)

A high-precision long distance time and frequency transfer mechanism used for clock offset determination or time synchronization between two stations.

U**Uniformly Most Powerful Invariant (UMPI) Test Statistic**

A test statistic that has uniformly the largest ► *power of all invariant statistics*

Universal Time (UT)

An irregular time scale based upon the rotation of the Earth. UT0 is a local time scale determined from observations taken at a single observing station. UT1 is UT0 corrected for the change in longitude of the observing station caused by polar motion. UT2 is UT1 corrected for seasonal variations. UT1 is proportional to the angle through which the Earth has rotated in space. The angular velocity of the Earth is proportional to the time rate-of-change of UT1.

Universal Time Coordinated (UTC)

Coordinated Universal Time is an atomic time aligned on the long-term on the Universal Time, i. e., the Earth's rotation. It is constructed by adding to the TAI, when needed, a leap second to keep the difference between UTC and UT less than 0.9 s. The difference between UTC and TAI is, therefore, always an integral number of seconds.

User Differential Range Error (UDRE)

A parameter broadcast by a ► *satellite-based augmentation system (SBAS)* to indicate the possible magnitude of the signal-in-space error for a specific satellite after applying the SBAS corrections. UDRE is determined from a broadcast 4-bit number, called the UDRE indicator or UDREI. A look up table is used to convert the indicator to a 1-sigma ► *overbound value* called the σ_{UDRE} . By tradition, UDRE itself is a 99.9% number or $3.29 \times \sigma_{UDRE}$.

User Equipment Error (UEE)

Contributions to the pseudorange measurement and modeling errors that relate to the user equipment (such as multipath and receiver noise). Atmospheric errors

such as residual tropospheric and ionospheric delays not taken into account by models or eliminated in a dual-frequency combination are also commonly attributed to the UEE.

User Equivalent Range Error (UERE)

The statistical error of the difference between observed and modeled pseudoranges that are used for computing a GNSS position solution. UERE is commonly split into contributions of the space and control segment (► *Signal-in-space Range Error, SISRE*) as well as contributions related to the user equipment and atmosphere (► *User Equipment Error, UEE*). Multiplication of UERE with the ► *dilution of precision* yields the statistical positioning error.

User Segment

The user equipment for tracking GNSS signals and for determining position, velocity, and time.

V**Vector Tracking Architecture**

A GNSS signal tracking architecture in which the local code and carrier replica generators are driven not by single-channel local feedback, as in a traditional *scalar* tracking architecture, but by the state estimate of a consolidated position, velocity, and timing estimator that takes in data from all active channels. A vector architecture benefits from the mutual information in code phase, carrier phase, and Doppler measurements between channels and can thus provide more accurate and robust tracking than a scalar architecture.

Vernal Equinox

The point of intersection of the ► *ecliptic* and the ► *equator*, at which the Sun crosses the Equator from south to north, during its yearly passage along the ecliptic. This currently occurs on about March 21 each year. Historically, the vernal equinox served as origin of the measurement of ecliptic longitude as well as right ascension in the ► *celestial coordinate system*.

Very Long Baseline Interferometry (VLBI)

A space geodetic technique utilizing microwave signals from extragalactic radio sources (quasars). Basically, the signal travel time difference between two radio telescopes is measured. VLBI is the only technique to determine ► *Universal Time UT1* and ► *nutation* parameters, and is used to realize the ► *International Celestial Reference System*.

VHF Data Broadcast (VDB)

The transmission of ► *Ground-based Augmentation System (GBAS)* differential corrections and integrity information using the ILS localizer's VHF frequency band (108–118 MHz) and a time division multiple access (TDMA) data format defined in the GBAS ► *ICD*.

VHF Omnidirectional Range (VOR)

An aircraft navigation system operating in the very-high frequency (VHF) band. VORs broadcast a VHF radio composite signal that allows airborne receiving equipment to derive the magnetic bearing

from the station to the aircraft. This line of position is called a *radial*.

Virtual Clock

A technique used by ► *Chinese Area Positioning System* to implement satellite navigation. With the satellite virtual atomic clocks, the time at which the signals are transmitted from the ground can be delayed into the time that the signals are transmitted from the satellites, and the pseudorange measuring can be fulfilled as in GPS.

Virtual Reference Station (VRS) Approach

An approach that presents the data of a network of multiple reference stations to the user or rover as if coming from a single reference station, referred to as the virtual reference station.

Viterbi Decoder

A device or software for decoding a navigation message encoded with a convolutional code for ► *forward error correction*. It builds on algorithms for optimal decoding first published by A.J. Viterbi in 1967.

W

Wavelength

The spatial separation between consecutive maxima or minima of an electromagnetic wave at a given instant of time.

Walker Constellation

A specific arrangement of multiple satellites in circular orbits around the Earth that enables good coverage and visibility conditions. Following J. G. Walker, the constellation geometry is described by the total number of satellites, the number of orbital planes, and the along-track shift of corresponding satellites in neighboring planes. All satellites within a plane exhibit identical spacing and the same applies for the ► *ascending nodes* of the individual orbital planes.

Warm Start

Activation of a GNSS receiver with prior information on the approximate time and user position, as well as the coarse orbit of GNSS satellites to speed up the signal search and acquisition.

W-Code

A low rate code used to encrypt the ► *P-code* of the GPS L1 and L2 signal. The product of the W- and P-codes yields the so-called ► *Y-code*.

Wet Delay

The wet component of the ► *slant total delay*.

Wide Area Augmentation System (WAAS)

A ► *satellite-based augmentation system (SBAS)* developed by the US Federal Aviation Administration (FAA) to provide horizontal and vertical navigation throughout North America. It has provided safety-of-life service since 2003.

Wide-lane Observable

A linear combination of carrier-phase observations on two frequencies that exhibits a large effective wavelength. It is formed as the difference of the two

carrier-phase observations expressed in cycles. For GPS L1 and L2 the wide-lane combination yields a wavelength of about 86 cm.

Wind-up

► *Phase wind-up*

World Geodetic System 1984 (WGS84)

A conventional terrestrial ► *reference frame* defined and maintained by the US Department of Defense. Nominally the ► *datum* for GPS ► *single point positioning*.

w-test Statistic

A ► *uniformly most powerful invariant (UMPI) test statistic* to test for the presence of one-dimensional biases. A special form of the w-test statistic is used in ► *data snooping* for the identification of observations contaminated with gross errors.

Y

Yaw-steering

The continuous control of a GNSS satellite's attitude around the Earth-pointing (yaw) axis to keep the solar panel axis perpendicular to the satellite–Sun direction.

Y-code

An encrypted version of the precise ranging code (P-code) transmitted by the GPS satellites on the L1 and L2 frequencies.

Z

Zenith

An imaginary point on the celestial sphere that is the projection of a local vertical direction. The astronomic zenith is the projection of the tangent to the local plumb line; the geodetic zenith is the projection of the local ellipsoid normal.

Zenith Total Delay

► *Slant Total Delay*

Zero-baseline

A setup of two or more GNSS receivers sharing a single antenna through a signal splitter.

Zero-tide System

A system specifically for the gravitational potential, in which all tidal effects except that of the permanent (mean) tidal deformation have been removed.

Z-tracking

An advanced technique for ► *semi-codeless tracking* of the GPS P(Y)-code on the L1 and L2 frequencies. The encryption signal bit is estimated separately in each frequency and fed to the other frequency to remove the encryption code from the signal. In this way, the code ranges and full wavelength L1 and L2 carrier phases are obtained. However, this method results in a signal-to-noise ratio degradation in comparison to the direct code correlation method.

Z-transformation

An integer preserving ambiguity transformation. A matrix is integer preserving if and only if all its

entries and those of its inverse are integer. Such transformations are used, e.g., in the ► *least-squares ambiguity decorrelation adjustment (LAMBDA)*

method, to re-parameterize ambiguities so that they can be estimated with higher precision and less correlation.

Subject Index

3 dB beam width 507
 3-D choke ring ground plane 521
 321 Euler angle sequence 785
³CAT-2 1183
 σ - μ -monitor 917

A

- Abel inversion 1144
- Abel transform 1123
- absolute calibration 1200
- absolute pseudorange 429
- absorption 167
- ACC 2.0 1006
- acceleration 940
 - drag 69
 - earth gravity 60
 - empirical 942, 948
 - gravitational 67
 - line-of-sight 935
 - nongravitational 941
 - radiation pressure 69
 - third-body 68
 - tidal 941
- accelerometer 816
 - pendulous 816
 - vibrating beam 816
- accumulation 472
- accuracy 841, 888, 910, 1013, 1275
 - limitation 1066
- acquisition 936, 1275
 - module architecture 408
 - performance 411
 - verification 409
- Adams–Bashforth–Moulton method 940
- adaptation 705
 - recursive 714
- additive white Gaussian noise (AWGN) 382, 404
- adjustment quality control 725
- ADS contract (ADS-C) 899
- advanced Bayesian estimation 837
- advanced GPS/GLONASS ASIC (AGGA) 937
- advanced RNP 887
- advisory circular (AC) 898
- aeronautical
 - information publication (AIP) 885
 - information regulation and control (AIRAC) 885
 - information services (AIS) 885
 - radio incorporated (ARINC) 884
 - radionavigation service (ARNS) 96, 470
- AFB 203
- afterslip 1086
- air force satellite control network (AFSCN) 204
- air navigation service provider (ANSP) 887
- air traffic control (ATC) 899
- air traffic services (ATS) 878
- aircraft
 - autonomous integrity monitoring (AAIM) 882
 - based augmentation system (ABAS) 340, 882
 - operator (AO) 896
- airport pseudolite (APL) 929, 1275
- airworthiness certification 897
- airy 33
- albedo 71, 991, 1275
- alert 1275
 - limit 889, 919, 1275
- algebraic reconstruction technique (ART) 1147
- ALGOS 156
- aliasing 380
- Allan deviation (ADEV) 121, 126, 243, 267, 291, 321, 1190, 1195, 1275
 - GNSS satellite 148
- Allan variance (AVAR) 124, 385, 813, 1275
 - modified 126
 - overlapping 126
- all-in-view (AV) 1193
- almanac 80, 229, 1275
- alternative BOC (AltBOC) 112, 116, 250, 403, 579, 592, 736, 1203, 1275
- alternative positioning navigation and timing (A-PNT) 900
- altimeter 942, 946
- ambiguity 1168, 1275
 - decorrelation 674
 - decorrelation number 671
 - dilution of precision (ADOP) 668, 1275
 - double-differenced 627
 - estimable PPP-RTK 628
 - fixing 15, 778, 1196
 - float solution 1275
 - function 1167
 - integer 15, 1286
 - spectrum 675
 - success rate 662, 678, 1275
- ambiguity resolution 665, 679, 771, 955, 996, 1012, 1275, 1286
 - full (FAR) 662, 678
 - partial (PAR) 662, 677, 1292
- ambiguity-fixed solution 1018, 1275
- AMCS 204
- amplitude modulation (AM) 122
- analog receiver architecture 366
- analog-to-digital conversion 380
- analog-to-digital converter (ADC) 365, 402, 551
- analysis center (AC) 730, 968, 970, 983, 1044
 - coordinator (ACC) 1002
- anechoic chamber 1275
- angular
 - momentum 1055
 - momentum vector 60
 - random walk (ARW) 817
 - rate vector 815
 - velocity 61
- anomaly
 - eccentric 61
 - mean 62
 - true 61
- antenna 1275
 - anti-jamming 517
 - array for radio occultation 519
 - bow-tie 513
 - calibration 574

- choke-ring 1278
- controlled reception pattern 517
- exchange (format) (ANTEX) 529, 572, 575, 727, 1004, 1227, 1276
- gain 507
- gain pattern 1275
- height error 689
- helix 510
- helix array 523
- microstrip 509
- patch 509
- patch array 524
- patch excited cup 517
- phase center 509, 989, 1275
- phase center variations 1112
- phase pattern 953
- phase-center offset 572
- pinwheel 513
- placement multipath mitigation 455
- quadrifilar helix 511
- reference point (ARP) 529, 573, 1275
- spiral 512
- swap 766
- temperature 1170
- testing 527
- thrust 991, 1276
- under test (AUT) 527
- anti-jamming technology 898
- anti-spoofing (A/S) 1124, 1276
- AO-40 934
- APEX 798
- apogee 61, 1276
- kick motor (AKM) 202, 1276
- Appleton–Hartree equation 168
- application specific integrated circuit (ASIC) 367, 426, 937
- approach
 - chart 893
 - procedure with vertical guidance (APV) 1276
 - with vertical guidance (APV) 889, 892
- appropriate means of compliance (AMC) 205, 898
- Aquarius 1177
- architecture evolution plan (AEP) 204
- arctangent discriminator 416
- area
 - correction parameters 1276
 - navigation (RNAV) 879, 1276
- argument of latitude 63, 72, 565
- argument of perigee 62, 565
- array antenna 375
- AS 17
- ascending node 62, 1276
 - secular drift 68
- assisted GNSS (A-GNSS) 847
- asthenosphere 1072
- Astronomical Institute of the University of Bern (AIUB) 951, 1000
- astronomical unit 1276
- atmosphere 1276
- atmospheric
 - density 941
 - excess phase 1121
 - loading 1094
 - parameter 974
 - propagation delay 728
 - signal delays 565
 - signature 1156
- atom clock ensemble in space (ACES) 136, 938
- atomic
 - clock 27, 1187, 1276
 - fountain 1276
 - fountain clock 135
 - frequency standard (AFS) 128, 235, 1276
 - time 27
 - time scale 1276
 - transition 128
- ATS route 878
- attenuation 167
 - factor 445
- attitude 781, 819, 1276
 - and orbit control system (AOCS) 266, 314
 - control 202
 - determination 787, 957
 - GNSS satellites 85
 - parameter 1276
 - parameterization 784
 - spacecraft 990
- augmentation
 - C/N_0 monitoring 493
 - distortion monitoring 493
 - information 1014
 - precorrelation structural power content analysis 493
- augmentation system
 - aircraft based (ABAS) 1275
 - BeiDou satellite-based (BDSBAS) 1277

- satellite-based (SBAS) 1296
- wide area (WAAS) 1301
- AUSPOS 1016
- authentication code 486
- autocorrelation 99, 402
- automated transfer vehicle (ATV) 955
- automatic
 - dependent surveillance (ADS) 892, 1276
 - dependent surveillance broadcast (ADS-B) 892, 899
 - direction finding (ADF) 878, 1276
 - flight control system (AFCS) 887
 - gain control (AGC) 372, 383, 480, 491
 - identification system (AIS) 863, 869
 - train control (ATC) 857
- autopilot (AP) 887
- availability 812, 841, 890, 1276
- avionics 884
- axial ratio 508
- Azimuth 1276

B

- Baikonur 237
- bandwidth 98, 507, 1276
 - antenna 507
 - signal modulation 1236
- bank angle 785
- Barker code (BC) 231
- Baro/VNAV 894
- barycentric
 - celestial reference system (BCRS) 148
- coordinate time (TCB) 26, 150
- dynamic time (TDB) 26
- system 1276
- base station 1014, 1277
- baseband 1276
- baseband signal 94, 406, 1277
- baseline 951, 953, 1277
 - error 625, 689
 - processing 1016
 - zero- 1301
- BDS-1 274
- BDS-2
 - positioning performance 295
 - service region 293
- beam steering 898

- beam width 507
 - beam-forming array antenna 519
 - BeiDou 139, 274, 881, 901, 1189, 1203, 1277
 - (Regional) Navigation Satellite System (BDS-2) 279
 - D1 navigation message 284
 - D2 navigation message 285
 - ephemeris parameters 285
 - ionospheric correction grid 285
 - navigation principle 277
 - Navigation Satellite System (BDS) 186, 273, 297, 506, 760, 978
 - operational control system 288
 - orbit determination 278
 - RAFS 140
 - receiver 371
 - satellite 286
 - satellite-based augmentation system (BDSBAS) 289, 358
 - signals and services 281
 - system architecture 275
 - time (BDT) 31, 159, 278, 291
 - BeiDou/Compass receiver 395
 - bending angle profile 1122
 - Bernese GNSS Software 940, 1000
 - best linear unbiased estimation (BLUE) 641, 1148, 1277
 - best linear unbiased prediction (BLUP) 651, 1277
 - between-receiver difference 1277
 - between-receiver single difference 594
 - between-satellite difference 1277
 - between-satellite single difference (BSSD) 596
 - bias 690, 817, 1277
 - characteristic 453
 - DGNSS 754
 - differential code 996, 1213, 1280
 - differential ionospheric 755
 - differential tropospheric 757
 - frame 46, 1283
 - influential 690, 1286
 - inter-channel 1286
 - inter-frequency 997, 1287
 - inter-satellite type (ISTB) 1287
 - inter-system 997, 1287
 - minimal detectable (MDB) 698, 1290
 - phase 1292
 - pivot receiver 755
 - satellite position 755
 - testable 690, 1298
 - bias-to-noise ratio (BNR) 691, 701, 1277
 - influential 691, 701
 - testable 693, 701
 - big endian 1217
 - binary exchange (format) (BINEX) 1217
 - binary offset carrier (BOC) 18, 109, 208, 252, 324, 394, 401, 451, 476, 579, 937
 - modulation 1277
 - multiplexed 1290
 - binary phase-shift keying (BPSK) 18, 107, 227, 251, 324, 369, 403, 448, 579, 937, 1277
 - modulation 1277
 - signal 448
 - bistatic
 - radar 1172
 - scattering coefficient 1168
 - scattering differential coefficient 1169
 - bit error correction 427, 1277
 - bit synchronization 1277
 - bit/symbol synchronization 425
 - blanking 499
 - Block 1277
 - block control 858
 - Block I satellite 199
 - Block II/IIA satellite 200
 - Block IIA 934
 - Block IIF satellite 201
 - Block IIR 934
 - Block IIR satellite 201
 - Block IIR satellite time keeping system 143
 - blunder 1277
 - body frame 819
 - Boltzmann constant 1169
 - bootstrapped PMF 668
 - bootstrapping 955, 1286
 - boresight angle 1277
 - Bortz orientation vector 820
 - Bose–Chaudhuri–Hocquenghem (code) (BCH) 284
 - boundary layer 1277
 - bow-tie turnstile antenna 513
 - box-wing model 69, 73, 942, 1277
 - bps 215
 - Brewster angle 1180
 - broadcast
 - ephemeris 83, 1213, 1277
 - ephemeris data 214
 - ephemeris model 81
 - group delay (BGD) 254, 261, 1277
 - buffer-gas-cooled ion standards 138
 - bulk acoustic wave (BAW) 127
 - Bureau International de l'Heure (BIH) 27, 37, 155, 221, 290, 1048
 - Bureau International des Poids et Mesures (BIPM) 27, 135, 223, 264, 292, 1041, 1188
-
- ## C
-
- C/A 16, 205
 - C/A-code 1278
 - generator 228
 - power spectrum 477
 - C/N₀ monitoring 491
 - cable delay 1199
 - cadastral survey 1028
 - calibration 1197
 - antenna 989
 - California earthquake 1069
 - California real time network 1070
 - campaign GPS network 1067
 - Canadian Meteorological Centre (CMC) 176
 - cannonball model 942
 - canonical interference model 476
 - carrier 5, 1278
 - pseudorange 432
 - tracking Kalman filter 437
 - carrier phase 6, 431, 1278
 - ambiguity 1278
 - bias 578
 - correction 908
 - differential GNSS (CDGNSS) 479, 908
 - error 450
 - measurement 563
 - multipath error envelope 452
 - multipath measurement 463
 - noise variance 433
 - observation 724
 - tracking 145
 - wind-up 569
 - carrier-Doppler aided code tracking 421
 - carrier-power to
 - interference-and-thermal-noise ratio (CINR) 473
 - carrier-range 1019

- carrier-to-noise 464
 - density ratio 1278
 - power-density ratio C/N_0 491, 578
 - ratio 405
- category I 895
- Cayley representation 786
- C-Band 1278
- celestial
 - coordinates 1278
 - ephemeris pole 1278
 - intermediate origin (CIO) 28, 54
 - intermediate pole (CIP) 1057, 1278
 - reference frame (CRF) 44
 - reference system (CRS) 44
 - sphere 26, 34, 1278
- cell-of-origin (COO) 850
- cellular network positioning 850
- Center for Orbit Determination in Europe (CODE) 72, 188, 745, 990, 1058
- center frequency 507
- center-of-gravity (COG) 947
- center-of-mass (CoM) 85, 987
- center-of-network (CoN) 998
- central bureau 734
- central processing unit (CPU) 389, 407, 457
- central synchronizer 1278
- Centre National d'Études Spatiales (CNES) 1001
- ceramic filter 378
- certification 897, 1278
- cesium 133 atom 26
- cesium beam frequency standard 130, 1278
- cesium clock
 - GLONASS 141
 - GPS 141
- CGGTTS time transfer standard 1192
- CHAMP 941, 959, 1144, 1182
- Chandler period 37, 987, 1278
- Chandler wobble 52, 988, 1057
- channel number 226
- Chapman layer function 178
- Chapman profile 1278
- China Geodetic Coordinate System (CGCS) 286
- China Geodetic Coordinate System 2000 (CGCS2000) 608
- Chinese Area Positioning System (CAPS) 298, 1278
 - concept 298
 - positioning 300
 - signal frequencies 299
 - system architecture 298
- chip scale atomic clock (CSAC) 137, 386
- chip technology 386
- chipping rate 206
- chirp 485
- chi-square distribution 691
- choke ring 456, 520, 1278
 - antenna 573
 - ground plane 520
- Cholesky decomposition 641
- Christian Doppler 92
- circular error probable (CEP) 50, 847, 973, 1279
- circular T 157, 1191
- Civil Aviation Authority (CAA) 356
- civil navigation message (CNAV) 81, 215, 232, 258, 310, 425, 489
- CL 208
- clean-replica
 - GNSS-R receiver 1164
 - waveform 1165
- clear zone 926
- climate
 - model 1119
 - monitoring 1118, 1128
 - research 1118
- clock 122
 - atomic 27, 1276
 - composite 30
 - constraint 996
 - densification 1000
 - difference 145
 - ensemble 1279
 - monitoring 145
 - monitoring and comparison unit (CMCU) 143
 - offset 1279
 - parameter 996
 - reference 1004
 - reference signals BDS 997
 - reference signals Galileo 997
 - reference signals GPS 996
 - relativistic effect 148
 - RINEX format 1221
 - spectrum 148
 - stability 122
- steering 430
- virtual 1301
- clock offsets example 146
- CM 208
- code
 - ambiguity 424
 - bias 1279
 - C/A 342, 367, 1279
 - correlation 409
 - direction 410
 - division multiple access (CDMA) 17, 97, 226, 267, 329, 379, 577, 588, 606, 754, 1099, 1202, 1279
 - M- 1289
 - Neuman-Hofman 1291
 - overlay 1292
 - P- 1292
 - P(Y)- 1294
 - phase 1279
 - pseudorange 428
 - ranging 1295
 - secondary 1297
 - shift keying (CSK) 312, 403, 1279
 - tiered 1299
 - W- 369, 1301
 - Y- 1301
- code/carrier generator 419
- code-carrier divergence (CCD) 95, 602, 913, 914
- code-minus-carrier (CMC) 460
- coherence 1165
- coherent
 - adaptive sub-carrier modulation (CASM) 113
 - integration 406
 - integration method 409
 - integration time 1279
 - population trapping (CPT) 130, 136
- cold atom clock 136
- cold start 936, 1279
- collision avoidance 870
- colocation site 1047
- colored noise 431
- combination
 - ionosphere-free 1287
 - narrow-lane 586
 - wide-lane 586
- combined processing 436
- Comité Consultatif International des Radiocommunications (CCIR) 155, 187

- commensurability 74
 - GPS 75
 - commercial channel navigation (C/NAV) message 256
 - commercial service 250
 - commercial-off-the-shelf (COTS) 938
 - common clock model 626
 - common clocks positioning model
 - estimable parameter function 627
 - common view (CV) 1192
 - communications-based train control (CBTC) 857
 - compact RINEX 1213
 - comparator 414
 - compatibility 1279
 - complementary metal oxide semiconductor (CMOS) 367, 936
 - composite binary offset carrier (CBOC) 112, 251, 252, 407, 451
 - generation block diagram 253
 - composite clock 30
 - compression
 - Hatanaka 1211, 1285
 - conductive test 540, 545, 1279
 - confidence
 - level 643
 - region 643
 - coning 813
 - constant envelope signal 113
 - constellation
 - Galileo 248
 - constrained maximum success-rate (CMS) 681
 - test 1279
 - construction machinery automation 1031
 - construction survey 1030
 - Consultative Committee for Time and Frequency (CCTF) 1198
 - Consultative Committee of Time and Frequency (CCTF) 1192
 - conterminous United States (CONUS) 341, 909, 1279
 - threat model 924
 - continental
 - drift 1279
 - hydrology loading 1094
 - US (CONUS) 878
 - continuity 812, 841, 890, 910, 1279
 - continuous GNSS network 1068
 - continuously operating reference station (CORS) 35, 311, 461, 650, 741, 761, 922, 1020
 - control
 - point 1279
 - segment (CS) 197, 203, 278, 983, 1279
 - surveys 1027
 - control segment (CS) 16
 - controlled flight into terrain (CFIT) 893, 1279
 - controlled radiation pattern antenna (CRPA) 367, 517
 - controller pilot data-link communications (CPDLC) 899
 - conventional navigation 878
 - conventional terrestrial pole (CTP) 290
 - convergence time 635
 - conversion gain 481
 - convex impedance ground plane 521
 - convolution 472
 - cooperative intelligent transport systems (C-ITS) 853
 - cooperative positioning (CP) 856
 - coordinate
 - celestial 33
 - time 149, 1279
 - coordinate system
 - Earth-centered inertial (ECI) 610
 - Earth-centered-Earth-fixed (ECEF) 610
 - local 611
 - local topocentric 64
 - Coordinated Universal Time (UTC) 29, 121, 198, 223, 251, 278, 319, 352, 398, 755, 927, 971, 1117, 1188, 1279
 - coordinates
 - Cartesian 31
 - celestial 1278
 - equatorial 1283
 - geocentric 1284
 - geodetic 32, 1284
 - geographic 1284
 - local 33, 1289
 - perifocal 1292
 - spherical 31
 - Coriolis acceleration 821
 - coronal mass ejection 922
 - correction models 985
 - correlation 472, 1165, 1279
 - cross- 1280
 - correlator 415, 1280
 - model 404
 - spacing 1280
 - coseismic 1280
 - displacement 1076, 1088
 - COSMIC 959
 - COSMIC/Formosat-3 1144
 - Cosmicheskaya Sistema Poiska Avariynyh Sudo (space system for search of distress vessels and airplanes) (COSPAS) 871
 - SARSAT 236, 250, 267, 871, 1280
 - Costas
 - loop 474, 475, 1280
 - PLL 416, 432
 - coupling
 - deep 826
 - course deviation indicator (CDI) 888
 - covariance 1280
 - matrix 944
 - Cramer Rao lower bound (CRLB) 104, 1280
 - CRCS-PPP 1016
 - CRISTA-SPAS 798, 957
 - critical inclination 68
 - cross plate reflector ground plane 522
 - cross-correlation 1164, 1280
 - protection 408
 - Crustal Dynamics Data Information System (CDDIS) 971
 - cryogenically cooled sapphire-loaded ruby oscillator 137
 - cryosphere 1177
 - cryptographic spoofing detection 495
 - crystal oscillator 376, 384
 - cut-off angle 995
 - cut-off elevation angle 772
 - cycle slip 433, 475, 484, 764, 1280
 - carrier-phase 689, 716
 - influential 691
 - cyclic redundancy check (CRC) 215, 232, 257, 311, 352, 392, 427, 895, 918, 1215
 - CYGNSS 1183
-
- ## D
-
- data 117, 885
 - assimilation 1141
 - bit 1280
 - bit transition 412, 416
 - bit/symbol demodulation 426

- center (DC) 968, 970
- channel 1280
- demodulation 424, 1280
- editing 992
- extract 428
- format 1004
- integrity 884
- quality monitoring (DQM) 913
- snooping 706
- symbol 1280
- datum 1280
- definition 998
- geodetic 34
- Davenport matrix 788
- day
 - Julian 26
 - solar 25, 28, 1297
- day boundary
 - discontinuity 79, 1196
 - jump 1195
- de Sitter 69
- dead reckoning 804
- Debye length 167
- decision altitude 1280
- decision height (DH) 883
- declination 33, 1280
- decorrelating Z-transformation 671
- decoupled clock model (DCM) 740
- dedicated short range communication (DSRC) 853, 855
- deep coupling
 - ultratight 826
- Defense Mapping Agency (DMA) 368
- deflection of the vertical 1280
- deformation 1280
 - survey 1027
- delay
 - and phase compensation 1166
 - Doppler coordinates 1167
 - Doppler-map (DDM) 1177
 - group 1285
 - hydrostatic 1286
 - lock loop (DLL) 368, 413, 448, 561, 826, 936, 1280
 - map (DM) 1176
 - slant 1297
 - total 1297
 - wet 1301
- delta range measurement 832
- Denali fault earthquake 1081
- detection 705
 - global 713
 - local 712
 - recursive 712
- detrending
 - polynomial 147, 148
- Deutsches GeoForschungsZentrum (GFZ) 21, 176, 745, 956, 1001
- Deutsches Zentrum für Luft- und Raumfahrt (DLR) 21, 187, 227, 325, 370, 515, 980
- DFH-3 platform 286
- DIA procedure 705
- dielectric coefficient 167
- dielectric loaded quadrifilar helical antenna 514
- difference
 - between-receiver 1277
 - between-satellite 1277
 - double 1281
 - single 1297
 - triple 1299
- differencing
 - between-receiver 624
 - between-satellite 631
 - double 632
 - single 631
 - triple 633
- differential
 - correction 444
 - correction generation 910
 - GPS (DGPS) 14, 708
 - positioning 14
- differential code bias (DCB) 576, 589, 613, 733, 734, 975, 1213, 1280
 - P1-C/A 614
 - receiver 616
 - satellite 613
- differential GNSS (DGNSS) 14, 466, 623, 753, 855, 907, 951, 952, 1281
 - correction latency 762
 - correction update rate 762
 - local 760
 - wide-area 760
- diffuse
 - reflection 446
 - scattering 1163
- digital
 - all-in-view receiver architecture 367
 - audio broadcast (DAB) 509
 - cesium beam frequency standard (DCFBS) 141
 - elevation model (DEM) 956
 - signal processing unit 388
 - signal processor (DSP) 378, 541, 578
 - video broadcasting (DVB) 509
- dilution of precision (DOP) 9, 259, 299, 322, 618, 1281
 - ambiguity 1275
 - geometric (GDOP) 618
 - horizontal (HDOP) 618
 - positioning (PDOP) 9, 184, 241, 263, 293, 333, 618, 835, 1293
 - vertical (VDOP) 618
- diode laser 136
- direct conversion 1281
- direction cosine 1281
 - matrix (DCM) 819
- discrete Fourier transform (DFT) 495
- discriminator 103, 414, 1281
- dispersive medium 566, 1281
- displacement 1281
- disposal orbit 1281
- distance measuring equipment (DME) 252, 341, 392, 484, 878, 1281
- distinct clock model 626
- distribution
 - τ - 710
 - beta 709
 - chi-squared 691
 - F - 709
 - normal 687, 707
 - t - 710
- DLL discriminator 417
- DLL discriminator function 448
- do-not-use flag 882
- Doppler 6, 1281
 - correlation 409
 - delay map 1281
 - direction 411
 - effect 93, 1281
 - measurement 563, 832
 - navigation 138
 - noise variance 433
 - observation 564
 - orbitography and radiopositioning integrated by satellite (DORIS) 945, 950, 968, 1040, 1222
 - range 1281
 - shift 5, 935, 1122, 1281
- Doppler/delay alignment 1281
- DORIS immediate orbit on-board determination (DIODE) 945
- double-delta correlator 458

double-difference (DD) 596, 670,
689, 723, 790, 952, 1000, 1281
– observation 464
downconversion 376, 1166, 1281
draconitic period 72, 1281
draconitic year 1046, 1281
drag 941
dry refractivity 170
DSSS 206
dual chip design on PCB 390
dual-frequency 1281
– GNSS pseudorange 724
dynamic
– displacement 1082
– time 26, 1281
– yaw steering 572
dynamo region 1156

E

early-minus-late correlator 1281
early-power minus late-power code
discriminator 417
Earth
– albedo 71
– gravitational potential 67
– oblateness 565, 1282
– observation (EO) 942
– orientation parameter (EOP) 46,
54, 973, 987, 996, 1054, 1282
– Parameter and Orbit System
(EPOS) 1001
– radiation pressure 1282
– rotation 25, 28, 1054
– rotation angle 1282
– rotation correction 155
– rotation parameter (ERP) 1004
Earth model
– fundamental parameter 222
– PZ-90 221
Earth-centered Earth-fixed (ECEF)
6, 148, 290, 352, 571, 787, 818,
985, 1282
– coordinate system 152
Earth-centered inertial (ECI) 148,
205, 985, 1282
– coordinate system 150
earthquake
– cycle 1073, 1282
– cycle deformation 1075
– magnitude 1078
– signature 1158
– warning 1083

Earth's center of mass (CoM) 733
East-North-Up (ENU) 787
east-north-up system 611
EC 17
eccentric anomaly 61, 564
eccentricity 60, 564
Echelle atomique libre (free atomic
scale) (EAL) 156
eclipse 1004
– transit 86
ecliptic 44, 1282
– obliquity 46
effective isotropic radiated power
(EIRP) 209, 1168, 1282
Efratom 139
EGM 205
eikonal 180
elastic block modeling 1077
elastic rebound hypothesis 1064
electric path length 171
electromagnetic
– band gap substrate 522
– roughness 446
– wave propagation 166
element, Keplerian 1288
elementary charge 1233
elevation 1282
– angle 785
– mask 995, 1282
ellipsoid 32, 1282
elliptic orbit 59
elongation 1282
Empirical CODE Orbit Model
(ECOM) 72, 995, 1282
end-fire mode 511
energy level 128
engineering surveying real-time
operation 1029
enhanced 911 (E911) directive 849
enhanced LORAN (eLORAN) 869
enhanced odometry 859
en-route navigation 891
envelope signal 94
EOPP 205
ephemeris 229, 1282
– broadcast 1213, 1277
– data 80
– JPL development 45
– solar system 941
– time (ET) 26, 1282
equation
– Kepler's 151
– of motion 939
– of origins 28

equator 27, 1282
– celestial 44
– radius 33
equatorial plasma bubble (EPB)
184
equator-S 934
equinox 44, 1283
– vernal 27, 28
error
– component 891
– systematic 1003
– variance matrix 653
eruption cycle 1090
estimation
– geometry of 690
– integer 661
– least-squares 639, 947, 1288
– recursive 644
– unbiased 641
estimator
– best linear unbiased 1277
– minimum mean square error
1290
Etalon 233
Euler
– angle 785, 819, 1283
– axis 820
– rotation theorem 820
– theorem 784
EUMETNET GNSS Water Vapour
Programme (E-GVAP) 1116
European Centre for Medium-Range
Weather Forecasts (ECMWF)
176, 729, 988, 1094, 1113, 1123
European Geostationary Navigation
Overlay Service (EGNOS) 19,
185, 248, 354, 762, 846, 883,
1141, 1283
– data access service (EDAS) 356
European Railway Traffic
Management System (ERTMS)
860
European Satellite Services Provider
(ESSP) 357
European Space Agency (ESA) 17,
243, 247, 370, 525, 541, 738,
1000, 1132
European Train Control System
(ETCS) 858
European TSO (ETSO) 897
evolved expendable launch vehicles
(EELV) 202
EWS 17
Excelitas 139

excess phase 1121
 executive monitoring 915
 expandable slot 198, 1283
 exponential relaxation 1088
 export regulation 938
 extended Kalman filter (EKF) 331,
 656, 824, 856, 954, 1283
 external Doppler-aided carrier
 tracking 421
 external reliability 701
 extra wide-lane 587, 669
 extreme atmospheric conditions
 758
 extreme ultraviolet (EUV) 12, 177,
 1152

F

F2 layer 1145
 F2 region 566
 fading 1283
 – depth 1149
 – frequency 447
 Falcon Gold 934
 false alarm 680
 – probability of 693, 707
 – region 700
 fast fading consideration 453
 fast Fourier transform (FFT) 389,
 408
 fault
 – detection 882, 890
 – detection and exclusion (FDE)
 882
 – monitoring 911
 Federal Communications
 Commission (FCC) 470, 849
 feedback 414
 fiber optic gyroscope (FOG) 815
 fiducial free 998
 field programmable gate array
 (FPGA) 373, 518, 541
 Figure-8 satellites 306
 filter 378, 650
 – extended Kalman (EKF) 331,
 656, 824, 856, 943, 954, 1283
 – fading-memory 713
 – finite-memory 713
 – information form 655
 – Kalman 436, 459, 653, 655, 710,
 824, 856, 994, 1288
 – SAW 378, 1298
 – unscented 943
 – variance form 655

final approach fix (FAF) 893
 final approach segment (FAS) 883
 – data block 895
 first-null beam width (FNBW) 507
 fishing fleet monitoring 870
 fix 879
 fixed failure rate ratio test (FFRT)
 680, 770, 1283
 fixed radiation pattern antenna
 (FRPA) 367, 517
 fixed solution 663, 679, 1283
 fixed-beam phased-array antenna
 456
 Flächenkorrekturparameter (FKP)
 776, 1283
 flattening 32, 1283
 fleet management 862
 flicker
 – drift (FLDR) 124
 – frequency (noise) (FLFR) 124
 – noise 1283
 – phase (noise) (FLPH) 124
 flight deck 884
 Flight Director (FD) 887
 flight management system (FMS)
 879, 884, 1283
 flight technical error (FTE) 891,
 1283
 FLL discriminator 416
 float solution 15, 663, 679, 793,
 1283
 floor noise model 1169
 FM 15
 FOH-YETE 914
 footprint 1283
 forecast 1283
 formation flying 951
 forward error correction (FEC)
 117, 215, 258, 311, 392, 425, 427,
 1283
 Fourier transform 124
 four-quadrant arctangent
 discriminator 416
 frame 229
 – bias 46, 1283
 – inertial 1286
 – international celestial reference
 1287
 – international terrestrial reference
 1287
 – orbital 86
 – reference 1296
 – synchronization 427, 1283

free navigation (F/NAV) message
 256
 free-space loss 523, 1283
 frequency 1283
 – comb 138
 – division multiple access (FDMA)
 17, 97, 226, 379, 577, 606, 736,
 754, 847, 997, 1202, 1284
 – drift 123
 – fractional 123
 – instantaneous 123
 – lock loop (FLL) 413, 1283
 – lock loop (FLL) discriminator
 416
 – normalized 123
 – offset 123
 – standard 122
 – unlock 475
 Fresnel
 – law of reflection 1174
 – radius 183
 – reflection coefficient 1174, 1177,
 1180
 – zone 183, 446
 Friis formula 379, 1284
 fringing 510
 front end (FE) 404, 1284
 front-to-back 509
 – ratio (FBR) 509
 full ambiguity resolution (FAR)
 677
 full operational capability (FOC)
 17, 248, 369, 731, 979
 functional model 688
 Fundamental Katalog 5 (FK5) 45
 future evolution 900

G

GaAs 201
 Gabor bandwidth 104, 250, 423
 gain pattern 507, 526
 – GPS IIR-M 524
 – M-shaped 523
 Galileo 247, 881, 901, 1189, 1203,
 1284
 – Control Centre (GCC) 270
 – FEC Encoder 258
 – Ground Segment 269
 – In-Orbit Validation Element
 (GIOVE) 238, 247, 524, 935
 – modulation details 252
 – passive hydrogen maser 145

- RAFS 140
- receiver 394
- spreading codes 254
- System Time (GST) 31, 159, 251, 263
- Terrestrial Reference Frame (GTRF) 261, 608
- Time Service Provider (GTSP) 263
- Galileo-GPS Time Offset 1284
- GAMIT-GLOBK 1001
- GANE 799
- GAS 17
- Gaussian distribution 1176
- Gauss–Jackson method 940
- Gauss–Newton iteration 612, 649
- GBAS 340, 457, 515, 556, 761, 854, 882, 890, 905, 1285
- Approach Service Type (GAST) 919
- error 926
- protection level 926
- GCOS Reference Upper Air Network (GRUAN) 1118
- GDOP 9
- general aviation 885
- generic
 - clock system 122
 - data/pilot multiplexing approach 435
 - GNSS signal 403
- geocenter motion 1045
- Geocentric
 - Celestial Reference System (GCRS) 44, 148
 - Coordinate Time (TCG) 26, 150, 1048
 - Terrestrial Reference System (GTRS) 148
- geodesy 1039
- geodetic
 - positioning 1064
 - survey applications 1023
- GEODYN 940, 947
- geodynamic 1063
- geographic information system (GIS) 514, 861, 1026, 1284
- geohazard monitoring 1043
- geoid 35, 1284
- geometric dilution of precision (GDOP) 9
- geometry screening 926
- geometry-free 585
- geometry-preserving 585
- GEONET 1070
- geostationary Earth orbit (GEO) 16, 61, 275, 305, 342, 398, 572, 770, 847, 934, 968, 1041, 1284
- geostationary satellite 347
- GEROS-ISS 1183
- GHOST 940, 947
- Gibbs vector 786
- GINS/DYNAMO 1001
- GIOVE-A 935
- glacial isostatic adjustment (GIA) 1064, 1069, 1095
- glistening zone 1173
- Global
 - Climate Observing System (GCOS) 1118
 - Ionospheric Scintillation Model (GISM) 184
- global
 - assimilation ionospheric model (GAIM) 1147
 - differential GPS (GDGPS) 854, 951
 - geodetic observing system (GGOS) 967, 1039, 1041
 - ionosphere map (GIM) 1141, 1222
 - ionospheric map (GIM) 568, 615, 728, 988
 - mapping function (GMF) 177, 569, 729, 986
 - navigation satellite system (GNSS) 339, 639
 - pressure and temperature (model) (GPT) 173, 569, 730, 761
 - pressure and temperature model (GPT2) 730
 - temperature trends 1129
- global navigation satellite system (GIPSY)
 - Inferred Positioning System and Orbit Analysis Simulation Software (OASIS) 940
- global navigation satellite system (GNSS) 3, 25, 59, 91, 121, 165, 197, 220, 250, 278, 305, 365, 401, 443, 469, 505, 535, 561, 583, 661, 687, 723, 753, 781, 811, 841, 905, 933, 967, 983, 1011, 1039, 1063, 1109, 1139, 1163, 1187, 1284
 - accuracy 1064
 - antenna 508
 - based reference frame 1050
 - disciplined oscillator 1189
 - Inferred Positioning System and Orbit Analysis Simulation Software (GIPSY) 1000
 - mixed-integer model 662
 - performance requirements 888
 - R model 1167
 - R receiver 1163
 - radio occultation 1144
 - receiver for atmospheric sounding 519, 1125
 - record and replay device 486
 - reflectometry (GNSS-R) 1163
 - satellite attitude modeling 570
 - satellite clock relativistic offset 154
 - signal diffraction 183
 - signal scattering 183
 - simulator 485, 1200
 - specific reference frame 1050
 - system time 606
 - time transfer technique 1191
 - timescale 158
 - tracking algorithms 1163
- Global Positioning System (GPS) 3, 30, 61, 96, 122, 181, 197, 219, 247, 281, 305, 340, 365, 401, 456, 473, 505, 536, 564, 586, 655, 661, 698, 723, 754, 782, 812, 843, 877, 905, 933, 934, 967, 983, 1011, 1039, 1064, 1110, 1140, 1164, 1188, 1284
 - Block IIR rubidium clock 145
 - C/A code L1 receiver 373
 - complementary service 309
 - constellation design 197
 - control segment 203
 - error source 343
 - III satellite 201
 - Klobuchar model 186
 - maneuver 75
 - P(Y) tracking 434
 - receiver application module 367
 - satellite 199
 - signal 207
 - signal legacy 205
 - signal overview 207
 - solar pressure model 73
 - Time (GPST) 30, 159, 319
 - Week 1285
- Global'naya Navigatsionnaya Sputnikova Sistema (Russian Global Navigation Satellite System) (GLONASS) 5, 39, 61, 139, 219, 232, 267, 305, 356, 369,
 - inferred positioning system and orbit analysis simulation software (GIPSY) 1000
 - mixed-integer model 662
 - performance requirements 888
 - R model 1167
 - R receiver 1163
 - radio occultation 1144
 - receiver for atmospheric sounding 519, 1125
 - record and replay device 486
 - reflectometry (GNSS-R) 1163
 - satellite attitude modeling 570
 - satellite clock relativistic offset 154
 - signal diffraction 183
 - signal scattering 183
 - simulator 485, 1200
 - specific reference frame 1050
 - system time 606
 - time transfer technique 1191
 - timescale 158
 - tracking algorithms 1163

426, 505, 536, 564, 588, 723, 753,
846, 881, 901, 933, 968, 984,
1048, 1099, 1125, 1147, 1173,
1188, 1201, 1202, 1284

- adjacent channel numbers 767
- ambiguity resolution 767
- central synchronizer 224
- cesium clock 145
- channel number 226
- clock estimation 997
- constellation parameter 221
- FDMA signal 226
- ground segment site 239
- inter-channel bias 767
- interchannel phase bias 766
- K 232
- launch vehicle 237
- M 232
- receiver 369, 394
- signal 225
- System Time (GLST) 31, 223
- time 31, 160

GOCE 941

Gold code 101, 228, 477, 1284

GPGGA 1208

GPS/MET 519, 799, 1144

GPS-aided GEO Augmented
Navigation (GAGAN) 19, 354,
762, 847, 883, 1141, 1284

GPS-to-Galileo time offset (GGTO)
263, 394

- dissemination 264

GRACE 517, 941, 956, 1091

gradient

- ionospheric 1287
- TEC 1143
- wedge 923

Gram–Charlier distribution 1176

GRAPHIC 590, 735, 945

- based PPP 735

graveyard orbit 249

gravitational

- coefficient 1233
- potential 35, 67, 940

gravity

- field 940
- wave 1130

Gravity Probe One 141

Green’s equivalent layer 1092

Greenwich

- apparent sidereal time (GAST)
28
- hour angle 1285
- Mean Time (GMT) 1285

- meridian 28, 37
- noon 27

grid ionospheric vertical error
(GIVE) 289, 343, 1285

Ground Control Segment (GCS)
17, 269

ground deformation surveys 1024

ground mission segment (GMS)
269, 370

ground monitoring station (GMS)
356

ground plane 509, 517, 1285

- convex 521
- corrugated 520

ground segment 1285

ground station 1285

ground track 63

- repeat 74
- repeat period 1233

ground uplink station (GUS) 347

ground-based

- augmentation system (GBAS)
340, 457, 515, 556, 761, 854, 882,
883, 890, 905, 1285
- monitoring 1110
- regional augmentation system
(GRAS) 908

group

- and phase ionospheric calibration
(GRAPHIC) 590, 735
- velocity 95, 1285

group delay 525, 1285

- broadcast 1213, 1277
- distortion 379
- timing 1213, 1299

GRS80 33, 1284

g-sensitivity 424

gyroscope 804, 815

H

Hadamard

- GPS satellite deviation 147
- variance 126

half-bit method 412

half-cycle ambiguity 432

half-power beam width (HPBW)
507

Hamming code 229

hand-over word (HOW) 210, 1285

HARDISP 987

hardware calibration 1198

hardware delay 1199

harmonic 1285

harmonics 484

Harvard–Smithsonian Center of
Astrophysics (CfA) 176

Hatanaka 1285

- compression 1211

Hatch filter 602

hazardous

- missed detection (HMD) 688
- probability 691
- region 700

heading 785

- sensor 869

health status 991

Heaviside function 107

height

- above threshold (HAT) 889
- ellipsoidal 35, 1282
- geoid 35, 1284
- system 1043

helicopter 885

helix 510

- antenna 374, 510
- quadrifilar 511

Helmert

- block method 647
- transformation 37, 40, 222, 998,
1285

high

- impedance ground plane (HIGP)
antenna 465
- resolution correlator 458
- sensitivity GNSS 847

higher-order ionospheric delay
correction 728

highly elliptical orbit (HEO) 398,
934

Hipparcos 46

homodyne down-conversion 378

horizon 1285

horizontal

- alert limit (HAL) 888
- dilution of precision (HDOP) 9,
249, 296, 825
- ionospheric gradient 755
- system 1285
- tropospheric gradient 758

hot start 1285

hour angle 1286

hydrogen maser 1194, 1286

- active 133
- frequency standard 132
- passive 134
- Q-enhanced 141

-
- hydrographic surveying 1033
 - hydrostatic troposphere delay 757
 - hydroxyl (OH) emission 227
 - hyperfine level 26
 - cesium 130
 - hydrogen 132
 - rubidium 129
 - hypothesis
 - alternative 689
 - null 689
 - signal acquisition 406
 - testing 692, 1286
-
- I
-
- ice sheet 1098
 - ICG 104
 - ICS 204
 - ID 210
 - identification 705
 - global 714
 - local 713
 - recursive 713
 - IERS reference meridian (IRM) 290
 - IERS2010 convention 732
 - IGb08 998
 - IGEX 220
 - IGLOS 220
 - IGS 13
 - antenna model 530
 - Central Bureau 969
 - clock combination 1003
 - final products 1002
 - multi-GNSS experiment (MGEX) 734
 - orbit accuracy 78
 - orbit combination 1002
 - organization 968
 - products 972
 - radiation pressure model 72
 - Real-Time Service (RTS) 976
 - reference frame 1050
 - spacecraft axes convention 85
 - structure 969
 - working group 971
 - IGS08 998
 - IIR 103
 - M satellite 201
 - image theory 445
 - impedance 508
 - inclination 62
 - least-squares (ILS) mixed 673
 - least-squares (ILS) optimality 674
 - mapping 663
 - recovery clock (IRC) 740
 - rounding (IR) 662, 666, 669, 1286
 - inclined geo-synchronous orbit (IGSO) 16, 61, 275, 305, 531, 564, 770, 979, 1286
 - Indian master control center (INMCC) 357
 - Indian navigation land uplink station (INLUS) 357
 - Indian reference stations (INRES) 356
 - Indian Regional Navigation Satellite System (IRNSS) 17, 31, 61, 140, 279, 305, 506, 731, 881, 978
 - constellation 322
 - RAFS 140
 - SPS code generator 326
 - inertia tensor 1055
 - inertial 1286
 - measurement unit (IMU) 497, 552, 799, 812, 852, 959, 1286
 - sensor 815
 - inertial navigation system (INS) 799, 812, 858, 959
 - computation 813
 - error state 814
 - vertical channel instability 823
 - influential bias 688
 - infrastructure data collection 861
 - initial approach fix (IAF) 893
 - initialization 654
 - in-orbit test (IOT) 260
 - in-orbit validation (IOV) 17, 74, 187, 247, 370, 524, 572, 731, 979
 - in-phase (*I*) 1286
 - component 94, 405
 - in-phase/quadrature (I/Q) 959
 - instantaneous impact point (IIP) 958
 - Institute of Electrical and Electronics Engineers (IEEE) 123
 - Institute of Space Device Engineering (ISDE) 369
 - instrument landing system (ILS) 340, 883, 906
 - instrumental delay 1195, 1199
 - integer
 - ambiguity 1286
 - ambiguity resolution 766
 - ambiguity search 674
 - aperture (IA) 679
 - bootstrapping (IB) 662, 667, 1286
 - cycle ambiguity 432
 - least-squares (ILS) 662, 673, 793, 1286
 - integrated water vapor (IWV) 1110
 - annual components 1119
 - diurnal components 1119
 - uncertainty 1115
 - integrity 812, 841, 877, 881, 882, 888, 910, 1286
 - beacon landing system (IBLS) 928
 - message 762
 - monitor testbed (IMT) 911
 - navigation (I/NAV) message 256
 - intelligent transport system (ITS) 842, 853
 - intensity 507
 - optical pumping (IOP) 130
 - intentional interference 485
 - interchannel bias (ICB) 606, 754
 - interface control document (ICD) 80, 220, 252, 279, 311, 371, 428, 578, 613, 881, 908, 1286
 - interference 469, 900, 1286
 - detection 491
 - mitigation 498
 - pattern 1182
 - rejection spread spectrum signal 98
 - spectrally 498, 499
 - temporally dense 499
 - temporally sparse 498
 - interferometric
 - GNSS-R 1165
 - GNSS-R receiver 1164, 1166
 - GNSS-R waveform 1165
 - noise factor 1170
 - synthetic aperture radar (InSAR) 661, 1088
 - interfrequency bias (IFB) 613, 754
 - interfrequency-channel bias (IFCB) 736
 - interleaving 427
 - intermediate fix (IF) 893
 - intermediate frequency (IF) 366, 402, 535, 536, 1287
 - intermediate origin, terrestrial 53
 - intermediate-lane combination 586
 - intermodulation product 114, 115
 - internal reliability 701

- International Association of Geodesy (IAG) 36, 1040
- International Association of Marine Aids to Navigation and Lighthouse Authorities (IALA) 867
- International Astronomical Union (IAU) 26, 47, 148, 1048
- International Atomic Time (TAI) 26, 27, 148, 332, 1192, 1287
- International Celestial Reference Frame (ICRF) 45, 973, 1287
- International Celestial Reference System (ICRS) 37, 45, 1287
- International Civil Aviation Organization (ICAO) 219, 358, 880, 889, 909
- International Earth Rotation and Reference Systems Service (IERS) 29, 37, 157, 172, 221, 290, 724, 970, 1024, 1041
- International GNSS Monitoring and Assessment (IGMA) 981
- International GNSS Service (IGS) 13, 73, 188, 222, 261, 313, 529, 573, 647, 724, 754, 868, 951, 967, 983, 1012, 1015, 1041, 1064, 1071, 1119, 1143, 1172, 1188, 1207
- formats 1218
- International Hydrographic Organization (IHO) 1034
- International Latitude Service (ILS) 36
- International Maritime Organization (IMO) 863
- international reference ionosphere (IRI) 187, 1148
- model 188
- international reference pole (IRP) 37, 290
- International Space Station (ISS) 799, 938, 1132, 1183
- International Telecommunication Union (ITU) 29, 96, 155, 227, 259, 274, 323, 470, 506
- International Terrestrial Reference Frame (ITRF) 6, 35, 37, 63, 216, 223, 261, 732, 968, 973, 984, 1015, 1022, 1039, 1044, 1048, 1064, 1220, 1287
- International Terrestrial Reference System (ITRS) 37, 148, 222, 290, 1039, 1048, 1287
- International Union of Geodesy and Geophysics (IUGG) 36, 148, 1040
- International Union of Radio Science (URSI) 155, 188
- interoperability 104, 1287
- interplex 113, 1287
- interpolation
- bilinear 1226
- clock 1220, 1222
- Flächenkorrekturparameter (FKP) 776, 777, 1283
- ionospheric 629
- Kriging 629
- orbit 1220
- SP3 1220
- VTEC 1225, 1226
- inter-satellite laser navigation and communication system 235
- inter-satellite-type bias, differential 770
- interseismic 1287
- period 1076
- intersignal correction (ISC) 577, 614, 735, 996, 1213, 1287
- intersystem bias (ISB) 577, 583, 606, 608, 617
- differential 768
- ionosphere-free 618
- intrack airport pseudolite 929
- inverted-F antenna (IFA) 514
- ionization 936
- ionosphere 177, 1287
- combination 589
- correction 189
- correction model 1287
- delay 566, 628, 988
- exchange (format) (IONEX) 728, 968, 1192, 1222
- higher order 988
- perturbation 1153, 1287
- refraction 1288
- ionosphere model
- Klobuchar 761
- NeQuick 761
- ionosphere-free 585
- coefficient 616
- combination 590, 613, 1192
- ionospheric
- delay 566, 627, 988
- disturbance flag 261
- divergence 460
- effect 1111
- gradient 689
- gradient monitoring (IGM) 926
- grid point (IGP) 347, 1287
- model 184
- path delay 944
- path delay, differential 953
- perturbation 1153, 1287
- pierce point (IPP) 343, 567, 1288
- radio occultation (IRO) 1145
- spatial decorrelation 922
- threat model 1152
- isodelay line 1167
- iso-Doppler lines 1167
- isostatic adjustment 1045
- issue-of-data (IOD) 262
- clock (IODC) 213
- ephemeris (IODE) 212
-
- ## J
- J2000 1288
- J_2 -correction 565
- Jacobian 944
- jamming 517
- Japan Aerospace Exploration Agency (JAXA) 306, 524
- Japanese Civil Aviation Bureau (JCAB) 356
- Japanese Geodetic System (JGS) 608
- Jason 950
- Jet Propulsion Laboratory (JPL) 45, 73, 137, 205, 742, 937, 990, 1000, 1084, 1124, 1147
- Joint Aviation Authorities (JAA) 880
- Joint Precision Approach and Landing System (JPALS) 898
- JPL 205
- Julian day/date (JD) 27
- number 1288
-
- ## K
- Kalman filter (KF) 436, 459, 653, 655, 824, 856, 994
- extended (EKF) 331, 656, 824, 856, 954, 1283
- unscented 837
- Kasami sequence 231
- Keplerian
- element 1288
- orbit 1288

Kepler's
 – equation 62, 151
 – law 59
 kinematic
 – GNSS positioning 1017
 – PPP solution 743
 kinematics 1014
 Klobuchar model 12, 945, 1288
 – RINEX navigation file 1213
 Korean Augmentation Satellite System (KASS) 358
 Kriging interpolation 629

L

L1 205
 – -SAIF signal 310
 L2C signal generation 208
 LAGEOS 941
 Lagrange
 – interpolation 1220
 – polynomial 1220
 land based applications 842
 land surveying operations 1024
 land transport applications 843
 lane level accuracy 843
 Laplace
 – plane 75
 – vector 61
 laser
 – cooling 136
 – residual 79
 – retro reflector (LRR) 266
 – retro-reflector array (LRA) 235, 287
 – time transfer (LTT) 288, 1288
 laser-cooled microwave standards 138
 last glacial maximum 1073
 latency 624, 805, 1288
 latitude 821
 launch and early orbit phase (LEOP) 238, 266
 launch vehicle 237
 L-band 1288
 leap second 29, 157, 1213, 1288
 Lear model 945
 least-squares
 – adjustment 993
 – ambiguity decorrelation adjustment (LAMBDA) 15, 587, 628, 676, 766, 793, 868, 955, 1288
 – batch 644

– constrained 647
 – estimation 8, 1288
 – geometry of 640
 – integer 1286
 – nonlinear 648
 – principle 639
 – rank-defect 647
 – recursive 645
 – weighted 640
 left-hand circular polarized (LHCP) 373, 508, 1171
 Legacy Accuracy Improvement Initiative (L-AII) 204
 legacy navigation (LNAV) 893
 – message 81, 210, 320
 – subframe 1 211
 – subframe 2–3 212
 – subframe 4–5 215
 legacy navigation message (LNAV)
 – error detection 211
 Legendre polynomial 941
 length-of-day 28, 973, 1055, 1288
 – variations 1056
 lense-thirring effect 69
 level crossing protection 861
 lever arm 805, 831
 LEX signal 311
 light detection and ranging (LIDAR) 838, 942
 light-time 1289
 line quality factor 128, 1289
 line replaceable unit (LRU) 884
 linear
 – feedback shift register (LFSR) 100, 228, 1289
 – ion trap standard (LITS) 142
 – quadratic Gaussian (LQG) 436
 – velocity vector 1073
 line-of-sight (LOS) 106, 180, 443, 519, 538, 1289
 – vector 609, 1289
 link budget 1289
 lithosphere 1072
 little endian 1217
 little ice age 1095
 Lloyd's mirror 1164
 LNAV/VNAV 893
 loading 39, 1289
 – displacement 1091
 – model 1093
 – tidal atmospheric 988
 – tidal ocean 987

local
 – area augmentation system (LAAS) 906
 – area differential GNSS (LADGNSS) 906, 907
 – area differential GPS (LDGPS) 898
 – coordinates 1289
 – oscillator 128
 – tie 1047
 localizer precision with vertical guidance (LPV) 341, 893
 LocataNet 930
 location based services (LBS) 842, 845
 LOD 985, 996
 logarithmic relaxation 1088
 London-moment 804
 long code (CL) 208, 392
 long range navigation (LORAN) 869
 longitude 821
 long-term secular reference frame 1045
 loop 1299
 – bandwidth 454
 – delay lock 1280
 – filter 414
 – frequency lock 1283
 – phase lock 1293
 – transient error 429
 loose (position domain) coupling 825
 loss
 – free-space 507
 – return 508
 Love number 43, 987, 1092
 Love's loading theory 1091
 low Earth orbit (LEO) 237, 576, 725, 799, 871, 934, 1109, 1144, 1176, 1289
 – satellite 726, 1109
 low visibility procedure 883
 low-noise amplifier (LNA) 366, 431, 481, 515, 577, 1289
 LSB 212

M

M/D ratio 450
 MA 97
 magneto-optical trap (MOT) 135

- Magnetosphere Multiscale Mission (MMS) 935
- Manchester
 - code 227
 - pulse 109
- mantle 1072
- mapping function 175, 1111, 1289
- NWP model result 176
- tropospheric 620
- mapping surveys 1028
- Mapping Temperature Test (MTT) 176
- maritime navigation 867
- Maser 1289
- Massachusetts Institute of Technology (MIT) 1001
- master
 - -auxiliary concept (MAC) 776, 1289
 - clock (MC) 291, 1289
 - control station (MCS) 17, 203, 275, 310, 356
 - station 1289
- matched filter 409, 1289
- matched-spectrum interference 478
- matrix
 - error variance 653
 - normal 640
 - reduced normal 647
 - test 714
 - transition 654
- maximum
 - a posteriori (MAP) 488, 824
 - coherence averaging interval 1170
 - likelihood estimation (MLE) 102, 417, 642, 1066, 1289
- Maxwell's equation 91, 166
- MC-LAMBDA 1290
- M-code 1289
- meaconing 487, 1289
- mean
 - anomaly 62
 - Earth ellipsoid 36
 - sea level 33, 35
 - sidereal time 63
 - solar time 1289
 - squared slope (MSS) 1172, 1176
 - tide system 43, 1289
 - time between cycle slips 475, 484
- meander sequence 227
- measurement
 - accuracy 1126
 - data age 833
 - model 993
- measurements quality monitoring (MQM) 914
- medium Earth orbit (MEO) 16, 61, 197, 221, 268, 274, 306, 526, 564, 770, 871, 934, 979, 1110, 1290
- medium-lane 587
- medium-scale traveling ionospheric disturbance (MSTID) 1157
- Melbourne–Wübbena combination 587
- MEOSAR 871
- meridian 28, 1290
 - Greenwich 37
- message
 - field range check 917
 - type 27 (MT27) 351
 - type 28 (MT28) 351
- MET 934
- metallic reflector ground plane 520
- Metop 519, 959
- MGEX Network 979
- Michibiki 307, 320
- micro-electromechanical system (MEMS) 127, 500, 804, 812, 852
- microstrip patch antenna 509
- microwave landing system (MLS) 340, 928
- midnight turn 86
- MIEV 925
- minimal detectable bias (MDB) 688, 698
- minimum
 - constraints (MC) 998, 1052
 - descent altitude 1290
 - mean penalty (MMP) 681, 1290
 - mean square error estimator 1290
 - obstacle clearance (MOC) 895
 - operational performance standards (MOPS) 11, 351, 880, 1290
- mining survey applications 1032
- Mir 955
- misalignment 817
- misalignment error 817
- missed approach point (MAPt) 893
- missed detection
 - hazardous 700
 - probability of 693, 707
- mixed mode 884
- mobile mapping system (MMS) 1032
- model
 - common clock 626
 - distinct clock 626
 - dynamic 654, 710
 - errors 689
 - GNSS attitude 790
 - ionosphere-fixed 625
 - ionosphere-float 625
 - ionosphere-weighted 625
 - Klobuchar 615
 - measurement 653, 710
 - Mogi 1088
 - NeQuick 615
 - PPP-RTK 627
 - Saastamoinen 615, 757
- model validation 687
 - batch 689
 - recursive 710
- moderate-length code (CM) 208, 392
- modified Julian day/date (MJD) 156
 - number 1290
- modulation 94, 107, 1290
 - binary offset carrier (BOC) 1277
 - binary phase shift keying 1277
 - multiplexed binary offset carrier 1290
 - time multiplexed binary offset carrier 1299
- Mogi model 1088
- moment magnitude 1078
- monitoring station (MS) 317
- Monte Carlo simulation 666, 676, 707
- Moon 28, 42, 47, 941
- Moore's law 387, 1290
- moving reference station DGNSS 755
- m-sequence 100
- multiaccess interference 478
- multibit quantization 479
- multi-constellation 901
- Multi-Function Satellite Augmentation System (MSAS) 19, 309, 354, 762, 846, 883, 1141
- Multi-function Satellite Augmentation System (MSAS) 1290
- Multifunction Transport Satellite (MTSAT) 356
- multifunctional chip 390
- multi-GNSS 978, 997
 - products 980
 - time transfer 1201

multi-instrument data analysis
 software (MIDAS) 1147
 multilateration (MLAT) 899
 multimodal PDF 665
 multipath 430, 443, 519, 759, 1163,
 1194, 1290
 – combination 591
 – combination, triple-frequency
 593
 – envelope 430
 – environment 444
 – error average 454
 – error envelope 106, 450
 – errors 579
 – impact, Doppler measurement
 466
 – limiting antenna (MLA) 926
 – measurement 459
 – mitigation 455, 519
 – ratio 509
 – rejection ratio (MPR) 509
 – relative delay 447
 – repeatability 462
 – suppression 375
 multiple receiver consistency check
 916
 multiplexed binary offset carrier
 (MBOC) 112, 209, 451, 1290
 multiplicative algebraic
 reconstruction technique (MART)
 1147
 multisignal message (MSM) 1215
 multistatic radar 1163
 multistep method 940
 mutual coherence 1164

N

nadir 1290
 NAGU 991
 narrow correlator 368, 450, 451
 narrowband
 – interference 476
 – radar ambiguity function 1168
 narrow-lane 1290
 – combination 586, 1196
 National Aeronautics and Space
 Administration (NASA) 3, 21,
 141, 317, 742, 937, 1047, 1132
 National Centers for Environmental
 Prediction (NCEP) 176
 National Geodetic Survey (NGS)
 530, 1001

National Geospatial-Intelligence
 Agency (NGA) 146, 204, 1051
 National Institute of Standards and
 Technology (NIST) 27, 135
 National Marine Electronics
 Association (NMEA) 397, 556,
 776, 1207
 – 0183 format 1207
 nautical mile (nmi) 879
 Naval Research Lab (NRL) 138
 navigation 1291
 – application 886
 – correction 833
 – data chain 885
 – data, RINEX 1213
 – frame 819
 – land Earth station (NLES) 356
 – message 5, 117, 613, 1291
 – message authentication (NMA)
 486, 1291
 – message correction table (NMCT)
 215
 – message structure 228
 – system error (NSE) 891, 911
 Navigation Package For Earth
 Orbiting Satellites (NAPEOS)
 940, 947, 1000
 Navigation with Indian Constellation
 (NavIC) 18, 61, 140, 305, 321,
 881
 near-field effect 1195
 near-real-time reconstruction 1148
 NeQuick 1291
 NeQuick model 13, 187
 – RINEX navigation file 1213
 NeQuickG 261
 network 760
 – RTK 1216, 1291
 – size 984, 985
 – time protocol (NTP) 1191
 networked transport of RTCM via
 Internet protocol (NTRIP) 15,
 762, 1215, 1291
 Neuman-Hofman (NH) code 232,
 282, 392, 1291
 neutral atmosphere 1110
 Newcomb 26, 47
 Newton's law of gravity 60
 next generation operational control
 segment of GPS (OCX) 204
 NiCd 200

Niell
 – mapping function (NMF) 176,
 569, 1111
 – Saastamoinen model 205
 noise 123, 1291
 – flicker 1283
 – temperatures 1169
 – term 1170
 – thermal 1298
 noiseless
 – GNSS-R waveform model 1169
 – waveform model 1169
 noncentrality parameter 691
 noncoherent integration 407
 noncut-off corrugated ground plane
 521
 nondirectional beacon (NDB) 878,
 1291
 no-net-rotation 37, 1291
 non-LOS signal 443
 nonorthogonality 817
 nonprecision approach (NPA) 879,
 889, 892, 1291
 nonrotating origin 28, 51
 nontidal ocean loading 1094
 nonzero-delay attack 488
 noon turn 86
 normal equation stacking 1000
 North American Datum (NAD) 14
 North celestial pole (NCP) 44
 North ecliptic pole (NEP) 44, 112
 North, East, Down (NED) 819
 notch filtering 499, 501
 notice
 – advisory 991
 – advisory to NAVSTAR users
 (NANU) 205, 914, 991, 1290
 – advisory to QZSS users (NAQU)
 320, 991
 – to airmen (NOTAM) 897, 1291
 NTCM model 187
 nuclear detection (payload)
 (NUDET) 201, 226
 null hypothesis 1291
 nulling attack 490
 numerical
 – integration 939
 – weather model (NWM) 172, 729
 – weather prediction (NWP) 176,
 1128
 numerically controlled oscillator
 (NCO) 402, 419, 563, 1291
 nutation 46, 985, 1054, 1058, 1291
 – transformation 49

Nyquist

- frequency 124
- sampling 381

O

oblateness 1282

- perturbation 67

obliquity 1291

- ecliptic 46
- factor 1291

observability 814

observation

- data 1209
- model, clock monitoring 145
- session 1014, 1291
- observation equation 428, 561
- carrier-phase 606
- differential 753
- double-differenced 763
- GNSS 606
- linearized 609
- multi-GNSS 607
- pseudorange 606

obstacle clearance 896

- surface (OCS) 924

occultation 1295

- mission 1124
- number 1125

ocean loading 39, 733, 1044

ocean wave spectrum 1176

OCX 204

odometry 859

on-board performance monitoring
and alerting (OPMA) 887

one-bit quantization 479

one-way carrier-phase technique
(OWCP) 146, 321Open Service (OS) 18, 250, 282,
943

open-loop tracking 1123, 1127

operational approval 898

operational control system (OCS)
204, 288, 912

optical

- clock 138
- comb 138
- frequency standard 1197
- molasses 135

optimal integer ambiguity test 681

OPUS 1016

OQPSK 115

orbit 1292

- accuracy 1003
- accuracy assessment 77
- and clock products 972
- arc length 993
- correction maneuver 75
- determination 1292
- elliptic 59
- error 625
- geostationary 64, 934, 1284
- graveyard 77
- highly elliptical 934
- inclined geosynchronous 63,
1286
- Keplerian 61, 151, 1288
- long-term-evolution 74
- low Earth 934, 1289
- medium altitude Earth 62, 934,
1290

- normal mode 86
- parameter estimation 995
- parameters 1233
- perturbations 66, 1292
- precision 1003
- repeat cycle 74
- repeat rate 61
- validation 78

orbit determination

- precise 1293
- reduced dynamic 938
- relative 952

orbit model

- almanac 80
- broadcast 79
- GLONASS 83
- Keplerian 61
- perturbed Keplerian 81

orbital

- parameters 1233
- period 61
- plane 60, 1292

orbital element

- Keplerian 63
- osculating 66

orbitography and synchronization
processing facility (OSPF) 262

orbits and clocks, precise 619, 622

orientation 781

origin

- celestial intermediate 28
- nonrotating 28, 51

original equipment manufacturer
(OEM) 366

orthonormality 784

oscillator 122, 1292

- jitter 424
- numerically controlled 1291

osculating element 66

outlier 1292

- code 689, 706, 716

- influential 691

oven controlled crystal oscillator
(OCXO) 127, 384, 488

overall model test 1292

overbound 1292

overlap comparison 950

overlay code 1292

oversampling 381

P

P 206

P wave 1078

P(Y)-code 1294

PAPR 113

parameter

- estimation 993

- GPS 198

parameterization 994

Parametry Zemli 1990 (PZ-90)
221, 608

parity bits 427

partial ambiguity resolution (PAR)
677, 1292

particle filter (PF) 837

parts-per-million 1292

Parus 4

passenger information systems (PIS)
863passive hydrogen maser (PHM)
142, 260

patch antenna array 524

path definition error (PDE) 891

P-code 1292

PCV map 948

PDOP 9, 618, 1293

pendulous accelerometer 816

performance

- based navigation (PBN) 877,
886, 1292

- requirements 877, 889

pericenter 60

perigee 61, 1292

- argument of 62

- secular drift 68

period

- draconitic 1281

- periodogram 148
- permanent GPS geodynamic array 1068
- permeability of vacuum 1233
- permittivity 510
 - of vacuum 1233
- personal
 - digital assistant (PDA) 386
 - navigation 841
 - privacy device (PPD) 470, 921
- perturbation
 - atmospheric drag 69
 - long periodic 66
 - radiation pressure 69
 - relativistic 69
 - resonant 74
 - secular 66, 68
 - short periodic 66
 - third-body 68, 941
- phase
 - ambiguity 627
 - center 1275
 - center variation (PCV) 376, 509, 574, 730, 948, 1050, 1112, 1227
 - error variance 474, 483
 - I receiver 366
 - integral 180
 - lock loop (PLL) 207, 372, 413, 474, 563, 579, 826, 936, 1083, 1127, 1190, 1293
 - modulation (PM) 113, 122
 - of flight 891
 - response 376
 - unlock 475, 484, 1293
 - velocity 92, 1293
 - wind-up 85, 731, 989, 1293
- phase center 1275
 - calibration 528
 - offset (PCO) 509, 529, 573, 730, 947, 989, 1227, 1292
 - offset (PCO), antenna effect 730
 - stability 509
 - variation 509, 515, 574, 947, 989, 1293
- phase-center
 - variation 575
- phased-array antenna 1293
- phase-range correction (PRC) 624, 1293
- phasor diagram 449
- physical constants parameters 1233
- piercing point 1141, 1142, 1288
- pilot
 - channel 117
 - signal 1293
- pincer defense 493
- pinwheel antenna 456, 513
- pitch angle 785
- pivot 1293
 - receiver 626
 - satellite 621, 763
- planar antenna 374
- planetary boundary layer (PBL) 1131
- planetary wave (PW) 1156
- plasma bubble 1150
- plasmasphere reconstruction 1145
- plate
 - boundary deformation 1073
 - boundary observatory (PBO) 1070, 1178
 - boundary zone 1072–1074, 1293
 - motion 1071, 1293
 - tectonic theory 1071
 - tectonics 1045
- Plesetsk 237
- PLL discriminator 416
- Poincare sphere 374
- point-in-space (PinS) 886
- polar motion 36, 973, 985, 1054, 1293
- polarimetric
 - phase interferometry (POPI) 1177
 - ratio 1177
- polarization 96, 508, 1171, 1293
 - linear 92
 - right hand circular 92
 - state 374
- pole
 - celestial ephemeris 50, 54, 1278
 - celestial intermediate 51, 54, 1278
 - north celestial 44
 - north ecliptic 44
 - rotation 1296
 - terrestrial 44
 - tide 733
- Pole, International Reference 37
- POLENET 1098
- polynomial
 - Lagrange 1220
 - Legendre 941
- pose 1293
- Position and Navigation Data Analysis (PANDA) 1001
- position dilution of precision (PDOP) 9, 184, 241, 263, 293, 333, 618, 835, 1293
- position error differential equation 829
- position, velocity and time (PVT) 256, 403, 470, 536
 - accuracy 471
- positioning 1293
 - accuracy 635
 - carrier-phase based 624, 764
 - DGNSS 624
 - global 633
 - kinematic 766, 938, 1288
 - local 634
 - multi-GNSS RTK 768
 - PPP-RTK 625
 - precise point (PPP) 613, 619
 - real-time kinematic 624, 763, 1295
 - regional 634
 - relative 1296
 - semikinematic 764, 1297
 - single point (SPP) 612, 615, 1297
 - static 1298
 - technologies 846
- positioning model
 - BeiDou (BDS), GLONASS 614
 - double-differenced 631, 763
 - GPS, Galileo, QZSS and IRNSS 613
 - point 612
 - relative 623
 - single point (SPP) 615
 - triple-differenced 633
 - undifferenced 626
- positioning solution
 - fixed 770
 - float 770
- positioning static 764
- positioning, navigation and timing (PNT) 3, 293, 305, 517, 863, 967, 1148
- postcorrelation FFT 409
- post-processed PPP service 742
- postseismic 1293
 - deformation 1085
- potential, tidal 42
- power
 - minimum received 1236
 - of the test 693
 - spectral density (PSD) 98, 404, 813

powerful near-band transmission 485
 PR 97
 – sequence 98
 preamble 432, 1293
 precession 28, 46, 985, 1293
 – geodesic 47
 – geodetic 69
 – luni-solar 47
 – planetary 47
 – transformation matrix 48
 precipitable water (PW) 1114
 precise
 – orbit determination 1293
 – orbit determination (POD) 320, 936, 946, 983, 992, 1054
 – positioning correction model 726
 – positioning service (PPS) 17, 216, 367, 507, 854, 1294
 – positioning techniques 1013
 precise point positioning (PPP) 13, 55, 313, 397, 479, 613, 619, 688, 723, 761, 778, 854, 938, 971, 1013, 1021, 1082, 1194, 1293
 – a priori correction model 727
 – consideration, GLONASS 736
 – data screening 726
 – dual-frequency 620, 622
 – dual-frequency model 621
 – editing 726
 – -model, between-satellite differenced 631
 – phase ambiguity fixing 739
 – positioning service 742
 – post-processed solution 741
 – real-time solution 742
 – redundancy 622
 – RTK 15, 625, 1294
 – RTK model 629, 630
 – single-frequency 620
 – single-frequency model 621
 – undifferenced 620
 precision 1294
 – approach (PA) 892, 1294
 predetection bandwidth 474
 prediction 650, 1294
 predictor, best linear unbiased 1277
 pre-elimination 948
 preprocessing 992
 primary spreading code 254
 printed circuit board (PCB) 386, 512
 PRISMA 956

probability
 – density function (PDF) 102, 642, 664, 692, 1169
 – mass function (PMF) 664, 797
 – of failure 680
 – of false alarm 693, 707, 1294
 – of hazardous missed detection 1294
 – of hazardous occurrence 1294
 – of missed detection 693, 707, 1294
 – of success 680
 – of successful fixing 680
 Procedures and Air Navigation Services (PANS) 880
 Procrustes
 – OPP 788
 – WOPP 789
 Program for the Adjustment of GPS Ephemerides (PAGES) 1001
 propagation
 – channel 539
 – delay 1111
 proper time 26, 149, 1294
 protection level 918, 1294
 proton 237
 PRS 18
 pseudolite 905, 928, 1294
 pseudo-random (PR) binary sequence 1294
 pseudo-random noise (PRN) 5, 146, 206, 227, 280, 309, 348, 408, 461, 699, 912, 1181, 1187, 1294
 – codes 325
 pseudo-random number sequence (PRN) 477
 pseudorange 6, 450, 1294
 – bias 576
 – correction (PRC) 624, 763, 907, 917, 1294
 – filtering 601
 – measurement 561, 832
 – noise variance 431
 – phase-adjusted 645
 – phase-smoothed 645
 – smoothing 601
 PSK 112
 public regulated service 250
 public-key cryptography 496
 pull-in range 416
 pull-in region 664, 1294
 – aperture 679
 – bootstrapping 668

 – least-squares 673, 676
 – rounding 666
 pulse
 – aperture correlator 458
 – blanking 393
 – clipping 393
 – per second (PPS) 545
 – shape 105
 – shape, band-limited 108
 – suppression 393
 pure PLL 416
p-value 693, 707, 1294
 PZ-90 39, 1292
 – transformation parameters 223

Q

QAM 94
 Q-enhanced hydrogen maser 141
 q-method 788
 quadrature
 – component 94, 405, 1294
 – phase-shift keying (QPSK) 114, 298, 403, 1294
 quadrifilar helix 1295
 – antenna (QHA) 511
 quality factor 127, 128
 – cesium clock 132
 – H-maser 134
 quantization 1295
 – effect 479
 – loss 380, 382
 quartz crystal oscillator 127, 1295
 quasi-instantaneous reference frame 1045
 Quasi-Zenith Satellite System (QZSS) 17, 31, 61, 140, 279, 305, 523, 565, 586, 731, 762, 846, 945, 970, 1051, 1127, 1188, 1295
 – constellation 307
 – control segment 317
 – signals 308
 quaternion 786, 820, 1295
 – estimator (QUEST) 789
 QZS-I satellite parameters 314

R

RADCAL 798, 957
 radiation
 – pattern 507
 – pressure 72, 942, 990, 1295
 – pressure, empirical 72

- radio
 - burst 1298
 - determination satellite service (RDSS) 274, 323, 1295
 - hologram 1181
 - navigation satellite service (RNSS) 274, 470, 506
 - occultation (RO) 519, 946, 947, 959, 1109, 1145, 1182, 1295
 - occultation (RO) measurements 1120
 - ranging 878
 - regulation 158
 - science 959
 - software defined (SDR) 1297
- radio frequency (RF) 127, 206, 251, 309, 365, 402, 456, 471, 506, 535, 908, 1165
 - chip technology 388
 - front-end 376
 - identification (RFID) 852
 - interference (RFI) 484, 913, 921
 - -processing 376
- Radio Technical Commission for Aeronautics (RTCA) 11, 393, 516, 880, 909
- Radio Technical Commission for Maritime Services (RTCM) 15, 313, 397, 742, 762, 868, 971, 1020, 1207
- radome 515, 1295
- rail applications 856
- Ramsey
 - cavity 130
 - pattern 131
- random
 - access memory (RAM) 372
 - walk drift (RWDR) 124
 - walk frequency (noise) (RWFR) 124
 - walk phase (noise) (RWPH) 124
- range-rate correction (RRC) 624, 763, 915
- ranging and integrity monitoring station (RIMS) 356
- ranging code 1295
- rank deficiency 763
- rapid-static positioning 1016
- rate-of-TEC (RoT) 1150
 - map 1143
- ratio
 - axial 508
 - bias-to-noise 1277
 - carrier-to-noise density 1278
 - front-to-back 509
 - test 679, 1295
- rationalisation of infrastructure 900
- ray tracing 1295
- Rayleigh-Taylor instability (RTI) 184
- read only memory (ROM) 372
- real-time
 - data stream 977
 - differential GNSS positioning 1019
 - navigation 942
 - nonlinear Bayesian estimation 838
 - service (RTS) 951, 971, 976, 1022
- Real-Time Analysis Center (RTAC) 976
- real-time kinematic (RTK) 15, 372, 573, 688, 741, 753, 774, 778, 854, 863, 908, 955, 1019, 1116, 1157, 1216, 1291, 1295
 - GLONASS 766
 - long-baseline 625
 - network (NRTK) 625, 774, 778
 - precise point positioning (PPP-RTK) 774, 778
 - short-baseline 624, 770
- real-time kinematic (RTK) positioning
 - convergence time 771, 773
 - epoch-by-epoch 772
 - long-baseline 772
 - multi-GNSS 773
 - precision 771, 773
- rebroadcasting test 541, 544
- received
 - power monitoring (RPM) 491
 - power monitoring (RPM), augmented 493
 - signal model 471
 - signal strength (RSS) 850
- receiver 1164
 - antenna calibration 530
 - architecture 365, 403, 457
 - autonomous integrity monitoring (RAIM) 341, 397, 882, 890, 897, 928, 1295
 - building block 372
 - clock 627, 1198
 - clock jump 429
 - code delay 627
 - geodetic-grade 1284
 - noise 578, 759
 - phase delay 627
 - single chip integration 390
- receiver autonomous integrity monitoring (RAIM) 897
- receiver independent exchange (format) (RINEX) 556, 577, 731, 776, 971, 1015, 1052, 1198, 1209, 1213, 1295
 - clock 1221
 - description 1209
- receiver-satellite
 - geometry 764
 - range 609
- record-and-playback system 536, 549, 1295
- reduced dynamic 938, 948, 995
- redundancy 639, 689, 1295
 - differenced models 633
 - number, local 698
- reference
 - ellipsoid 1296
 - frame 34, 608, 1296
 - frame station 974
 - frame, celestial 44
 - station 1014, 1296
 - station network 760
 - station, continuously operating (CORS) 35, 1279
 - station, virtual 1301
 - system 34, 1296
 - system, celestial 44
 - system, geodetic 1284
 - time scale 1188
- Reference Frame
 - International Celestial 45, 973, 1287
 - International Terrestrial 6, 37, 63, 216, 223, 261, 732, 968, 973, 984, 1015, 1022, 1039, 1044, 1048, 1064, 1220, 1287
- Reference System
 - International Celestial 37, 45, 1287
 - International Terrestrial 37, 148, 222, 290, 1039, 1048, 1287
- reflection coefficient 445
- reflectometry 518, 959, 1296
- reflector-backed monofilar antenna 526
- refraction 1296
- refractive index 565
- regional argumentation 1296
- regional navigation satellite system (RNSS) 16, 305

regulation 897
 relative
 – calibration 1200
 – position accuracy 1064
 – pseudorange 429
 relativistic
 – correction 286
 – effects 564
 relativistic clock correction 1221
 – periodic 154
 – rate offset 154
 relativity 148, 989
 relaxation model 1087
 reliability 841
 remote clock comparison 1191
 rendezvous 951
 replica generator 414
 required navigation performance
 (RNP) 864, 886
 – approach (RNP APCH) 887, 893
 – approval required (RNP AR) 887,
 895
 required navigation performance and
 special operational requirements
 study group (RNPSORSG) 886
 residuals
 – least-squares 696
 – normalized 697
 – predicted 644, 711
 – studentized 710
 resonance 74
 return loss 508
 reversion 900
 REX-II 799, 957
 RHCP 92
 Richter scale 1078
 right ascension 33, 1296
 – of ascending node (RAAN) 62,
 198, 248, 307
 right ascension of ascending node
 (RAAN) 62
 right-hand circular polarized (RHCP)
 373, 507, 569, 731, 1171
 rigid plate motion 1071
 ring laser gyroscope (RLG) 815
 RNAV approach 894
 road level accuracy 843
 ROCK model 990
 Rodrigues
 – modified vector 786
 – vector 786, 1296
 roll angle 785

root mean square (RMS) 77, 104,
 188, 209, 222, 259, 320, 418, 744,
 946, 1083, 1178
 root-sum-square (RSS) 696, 918
 rotation matrix 819
 roughness parameters 1176
 rounding 666
 – integer 666, 669, 1286
 – vectorial 666
 RS 19
 RTCM 776, 1296
 RTCM (format) 1296
 – message 762
 – multisignal messages 1215
 RTCM SC-104 (format) 1207,
 1215
 RTS Network 976
 rubidium atomic frequency standard
 (RAFS) 129, 140, 261, 287, 315,
 1296
 Runge–Kutta 939
 Runge–Lenz vector 61
 Russian Federal Space Agency
 (RFSA) 238

S

S wave 1078
 S(ingularity)-transformation 648
 S₄ index 424, 1148
 Saastamoinen hydrostatic delay
 model 173
 Safety of Life at Sea (SOLAS) 863
 Sagnac
 – correction 562, 1296
 – effect 152, 155, 804
 sample rate 1296
 sampling rate 381
 satellite
 – antenna calibration 530
 – clock 623, 627
 – clock offset 613
 – clock, dithering 754
 – code delay 627
 – failure 689
 – geostationary 1284
 – laser ranging (SLR) 3, 37, 38, 74,
 78, 233, 315, 942, 950, 967, 1003,
 1040, 1064, 1222, 1296
 – phase delay 627
 satellite-based augmentation system
 (SBAS) 11, 61, 185, 216, 282,
 311, 339, 427, 506, 515, 556, 761,
 846, 882, 890, 908, 970, 1141
 – architecture 345
 – integrity 349
 – message type 352
 – user algorithm 351
 S-band 1296
 scalar factor error 817
 scale factor (SF) 817
 scatterometry 518, 959, 1173
 Schuler oscillation 823
 Schwarzschild perturbation 69
 scintillation 424, 482, 912, 1296
 – index 483
 – measurement network 1151
 – monitoring 1148
 – receiver 1149
 sculling 813
 S-curve 103, 448
 – shaping 458
 sea
 – ice 1179
 – ice permittivity 1180
 – level monitoring 1052, 1053
 – surface altimetry 1173
 – surface permittivity 1177
 – surface roughness 1173
 – surface scatterometry 1175
 search and rescue (SAR) 201, 256,
 525, 870, 871, 947, 1296
 search halting 675
 search space 410
 search-and-expand algorithm 794
 search-and-shrink algorithm 795
 second 26, 1297
 secondary code 254, 412, 1297
 – synchronization 425
 secondary surveillance radar (SSR)
 899
 SECOR 4
 secular perturbation 66
 security code 486
 security code estimation and replay
 (SCER) 487
 – attack 487, 497
 security-enhanced GNSS signal
 486
 seismic wave 1082
 seismology 1078
 seismometer 1082
 selective availability (SA) 14, 216,
 225, 395, 754, 880, 898, 905, 938,
 1121, 1297

- selective availability anti-spoofing module (SAASM) 367
- semi-codeless tracking 1297
- semi-major axis 32, 59, 564, 1297
- semi-minor axis 59
- sensor fusion 855
- sensor-inherent noise 817
- service, standard positioning (SPS) 1298
- S-function 448
- SGLS 203
- shadowing 444
- Shapiro effect 564, 1297
- Shida number 43, 732, 987
- shielding chamber 1297
- shipboard relative GPS (SRGPS) 898
- shipping container tracking 872
- short-delay multipath 461
- SI second 26, 1191
- sidelobe 934
- sidereal
 - day 1297
 - filtering 459
 - time 27, 1297
- signal
 - deformation monitoring (SDM) 912
 - model 404
 - model, multipath 448
 - multipath 1112
 - power 434
 - search-acquisition 406
 - structure 1235
 - tracking 413
 - travel time 561
- signal-in-space (SIS) 216, 880
- signal-in-space range error (SISRE) 9, 10, 84, 241, 282, 333, 945, 1051, 1297
- signal-in-space receive and decode (SISRAD) 911
- signals of opportunity (SOO) 838
- signal-to-interference-plus-noise ratio (SINR) 98, 1297
- signal-to-noise ratio (SNR) 98, 129, 408, 464, 471, 848, 1065, 1121, 1164, 1170, 1297
- significance level 1297
- similarity transformation 37
- simulator 535
 - IF-level 541, 546, 1286
 - key requirements 541
 - measurement-level 541, 552, 1290
 - record and playback system 542, 549
 - RF-level 540, 543, 1296
 - types 540
- simultaneous location and mapping (SLAM) 838
- single event
 - latch-up (SEL) 936
 - update (SEU) 936
- single photon avalanche diode (SPAD) 288
- single point positioning (SPP) 13, 612, 754, 938, 1297
 - between-satellite differenced model 631
 - dual-frequency 616, 617
 - multiconstellation 617
 - redundancy 616
 - single-constellation 615
 - single-frequency 615, 617
- single-difference (SD) 596, 740, 790, 952, 1297
- single-element antenna 456
- single-frequency 735
 - point positioning (PP) 735
- single-input-single-output (SISO) 421
- site
 - coordinates 973
 - displacement 985
 - displacement effect 732
 - log 1228
- skew representation 786
- skyplot 65
- slant total
 - delay 1297
- electron content (STEC) 178, 567, 728, 1140
- slot number 221
- small atomic clock technology 136
- Smithsonian Astrophysical Observatory (SAO) 141
- smooth spectrum approximation 476
- smoothing 601, 657
 - fixed-interval 657
 - fixed-lag 657
 - fixed-point 657
- SMOS 1177
- SMS 17
- Snell's law 444, 446
- snow depth 1181
- software defined radio (SDR) 366, 518, 1297
- soil moisture 1179, 1181
- solar
 - maximum 482
 - radiation pressure (SRP) 942, 1298
 - radio burst (SRB) 481
- solar radiation pressure (SRP) 990
- solid earth tide 732
- solution
 - fixed 1283
 - float 1283
 - separation 695
- solution independent exchange (format) (SINEX) 974, 1002, 1048, 1222, 1298
 - bias 1222
 - troposphere 1222
- sounding rocket 958
- Soviet Geodetic System 221
- Soyuz 238
- Space
 - Integrated GPS/Inertial navigation system (SIGI) 958
 - Service Volume 935
 - Shuttle/SIR-C 1182
- space
 - based augmentation systems 850
 - environment 517
 - hydrogen maser (SHM) 142
 - rubidium atomic clock 139
 - segment 16, 197, 1298
 - vehicle (SV) 199
 - vehicle number (SVN) 77, 146, 199, 912, 1227, 1298
 - weather 1152, 1298
- space-qualified
 - cesium beam clock 140
 - hydrogen maser clock 141
- space-time
 - curvature 69, 562
 - interval 149
 - reference system 148
- spatial incoherence 1164
- specific force 816
- speckle noise 1165, 1171
- spectral
 - analysis 494
 - density 124
 - lines 476
- spectrum 93
- specular reflection 446
- speed of light 1233

spherical harmonic 35, 67, 940
 spheroid 32
 spiral antenna 512
 split-spectrum modulation 111
 spoofing 485, 517, 1298
 spreading code 94
 spread-spectrum processing gain 476
 spread-spectrum signal 97
 square root information filter 994
 squaring loss 408, 435, 474
 S-system 616, 628
 stability
 – clock 123
 – frequency 123
 standard positioning service (SPS) 16, 216, 311, 507, 846, 880, 881, 915, 943, 1298
 Standard Product 3 (format) (SP3) 727, 1004, 1022, 1218
 standards and recommended practices (SARPS) 880, 909
 standards of fundamental astronomy (SOFA) 55
 state
 – estimation 812, 824
 – space representation (SSR) 15, 977
 – space representation (SSR) messages 1216
 – transition matrix 944
 – vector 813, 939
 static
 – displacement 1079
 – positioning 1014
 – PPP solution 743
 station clock solution 745
 stochastic orbit parameter 1298
 Stoermer–Cowell method 940
 stop-&-go GNSS survey 1019
 straight line tangent point altitude (SLTA) 1122
 strapdown
 – algorithm (SDA) 818
 – inertial navigation 818
 stratosphere 168, 1298
 strobe correlator 457
 success rate 666
 – ambiguity 662, 664, 678
 – bootstrapping 668
 – least-squares 676
 – rounding 666
 success-rate bounds
 – bootstrapping 668

 – least-squares 677
 – rounding 666
 Sun 28, 42, 47, 941
 – elevation angle 72
 superframe 229
 superheterodyne down-conversion 378
 surface
 – acoustic wave (SAW) 127, 372, 430
 – acoustic wave (SAW) filter 378, 1298
 – deformation 1075
 – loading deformation 1091
 surface-to-mass ratio 941
 surplus satellite 198, 1298
 surveillance 899
 surveying 1011
 – cadastral 1278
 – construction 1279
 – hydrographic 1286
 SVD 788
 symmetric-key cryptography 496
 synthetic aperture radar (SAR) 943, 946, 1043
 system
 – barycentric 1276
 – frame 34
 – inertial 1286
 – mean tide 43, 1289
 – noise temperature 379
 – of observation equations, rank-deficient 763
 – reference 34, 1296
 – tide-free 43, 1299
 – time (ST) 159, 1157
 – time offset 607
 System for Differential Corrections and Monitoring (SDCM) 19, 240, 354, 847, 883
 system-on-a-chip (SoC) 372

T

tactical air navigation (system) (TACAN) 341, 392, 484
 tactical air navigation system (TACAN) 484
 TanDEM-X 956
 TDRSS augmentation service for satellites (TASS) 945
 technical standard order (TSO) 897
 tectonic plates 37, 39
 telemetry (word) (TLM) 210, 326
 telemetry, tracking, and commanding (TT&C) 262, 286, 314, 525
 temperature compensated crystal oscillator (TCXO) 128, 384, 411, 488
 temperature variation 1195
 temporal error variation 453
 terminal area 892
 Terralite XPS 929
 TerraSAR-X 517, 959
 terrestrial
 – dynamic time (TDT) 26
 – intermediate origin (TIO) 53
 – reference frame (TRF) 985, 998, 1040, 1047, 1094
 – reference system (TRS) 1047
 – time 26, 1298
 – time (TT) 26, 150
 test
 – ambiguity acceptance 679
 – conductive 1279
 – constrained maximum success-rate (CMS) 681, 1279
 – difference 679
 – fixed failure rate ratio 680, 1283
 – global overall model (GOM) 713
 – hypothesis 1286
 – local overall model (LOM) 712
 – minimum mean penalty (MMP) 681, 682, 1290
 – optimal integer ambiguity 681
 – overall model 1292
 – projector 679
 – ratio 679, 1295
 – rebroadcast 1295
 – uniformly most powerful invariant (UMPI) 693
 test statistic 1298
 – T_q 696
 – UMPI 695, 711
 – w - 697
 testable bias 688
 testing
 – global 711
 – local 711
 – procedure 705
 – recursive 711
 thermal-noise-equivalent approximation 471
 thin-shell ionospheric model 567
 threat model 922
 three-carrier ambiguity resolution (TCAR) 587

- tide 942, 1044, 1298
 - body 42
 - direct 68
 - ocean 42, 68
 - ocean pole 988
 - pole 987
 - solid Earth 39, 68, 987
- tide-free 987
 - system 43, 1299
- tiered code 102, 255, 1299
- tight (observable domain) coupling 826
- tightly combined processing 768
- TIMATION (Time Navigation) 139
- Time
 - Coordinated Universal (UTC) 29, 1279, 1300
 - International Atomic (TAI) 27, 1287
 - Universal (UT) 28, 1300
- time 25
 - atomic 27
 - barycentric coordinate 26
 - BeiDou 31
 - coordinate 149, 1279
 - division multiple access (TDMA) 97, 1299
 - dynamic 26
 - dynamical 1281
 - ephemeris 26, 1282
 - Galileo 31
 - geocentric coordinate 26
 - GLONASS 31, 160
 - GNSS 158
 - GPS 30, 216, 1284
 - interval counter (TIC) 1198
 - keeping system (TKS) 143
 - laboratory 1188
 - multiplexed binary offset carrier (TMBOC) 112, 1299
 - multiplexed binary offset carrier (TMBOC) modulation 1299
 - proper 26, 149, 1294
 - receiver 1197
 - scale 1299
 - sidereal 27, 1297
 - synchronization 424
 - terrestrial 26, 1298
 - transfer 1299
- time offset
 - GPS-to-Galileo (GGTO) 607
 - system 607, 608
- time system differences
 - RINEX navigation file 1213
- time-of-arrival (TOA) 850
- timescale atomic standard 135
- timescale, IGS 1004
- time-to-alert (TTA) 340, 888, 1299
- time-to-first-fix (TTFF) 386, 412, 849, 1299
- timing group delay (TGD) 286, 614, 735, 996, 1299
- tolerable error limit (TEL) 924
- tomography 1299
- TOPEX/Poseidon 934
- topographical surveys 1028
- TopSat 958
- total
 - ionization dose (TID) 936
 - station 1299
 - system error (TSE) 891
- total electron content (TEC) 12, 178, 519, 566, 584, 746, 756, 971, 988, 1139, 1140, 1222, 1299
 - change index (ROTI) 1150
 - gradient 1143
 - map 1141
 - measurement 1140
- tracking 1299
 - and data relay satellite system (TDRSS) 945
 - jitter 422
 - loop 6, 414, 936, 1299
 - loop error 423
 - loop model 414
 - network 984
 - performance 422
- traditional navigational 341
- train control 857
- trajectory model 939
- transformation
 - celestial/terrestrial 46
 - CIO method 54
 - decorrelating 670
 - equinox method 54
 - Helmert 37, 40
 - nutation 49
 - precession 48
 - reference frame 985
 - similarity 37
- transient
 - error 431
 - slip 1085
- transionospheric radio wave 179
- Transit (satellite) 4, 1033
- transit system 138
- transport rate 820
- traveling
 - ionospheric disturbance (TID) 1157, 1299
 - wave tube amplifier (TWTA) 287, 324, 1299
- triangular decomposition 668, 675
- triple-difference 598, 1299
- tropopause 168, 1129
- troposphere 168, 568, 1299
 - characteristic 168
 - delay estimation 174
 - empirical model 172
 - gradients 995
 - hydrostatic delay 988
 - mapping function 988
 - modeling 988
 - parameters 995
 - wave propagation 166
- tropospheric
 - delay 566
 - delay modeling 728
 - mapping function 758
 - refraction 170, 1300
 - zenith path delay 745
- true anomaly 61, 565
- Tsikada 4, 219
- TSO C-129 897
- tsunami warning 1083
- Tsyclon 219
- TT&C 201
- Tundra 306
- two way time transfer 1204
- two-body problem 60, 1300
- two-way satellite time and frequency transfer (TWSTFT) 159, 291, 315, 1300

U

- U-D 205
- UK Disaster Monitoring
 - Constellation (UK-DMC) 519, 1172, 1182
- UK TechDemoSat-1 (UK-TDS1) 1182
- ULS 204
- ultra stable oscillator (USO) 1121
- ultra-high frequency (UHF) 15, 201, 483, 523, 909, 1021
- ultra-wideband (UWB) 851
- uncalibrated phase delay (UPD) 740
- under sampling 381

uniformly most powerful invariant (UMPI) 693, 1300
 United States Geological Survey (USGS) 368, 1072
 United States Naval Observatory (USNO) 27, 121, 205, 319, 398, 1188
 Universal Time (UT) 26, 28, 922, 973, 1049, 1300
 – Coordinated (UTC) 198, 1188, 1279, 1300
 University NAVSTAR Consortium (UNAVCO) 21, 1024, 1074
 University of New Brunswick (UNB) 176, 729
 unmanned aerial vehicle (UAV) 800, 909
 unmanned ground vehicle (UGV) 804
 unmodeled bias 759
 unscented Kalman filter (UKF) 837
 UoSat 958
 – -12 799
 update, measurement (MU) 655, 710
 update, time (TU) 654, 710
 urban canyon 9
 US Federal Aviation Administration (FAA) 14, 216, 341, 397, 880, 909
 user differential range error (UDRE) 289, 343, 1300
 user equipment error (UEE) 9, 1300
 user equivalent range error (UERE) 9, 259, 333, 1300
 user ionospheric range error (UIRE) 353
 user ionospheric vertical error (UIVE) 353
 user range error (URE) 213, 295
 user segment 16, 1300
 UT1 985
 UTC dissemination 264

V

van Cittert–Zernike theorem 1165, 1170
 variance
 – Allan 124
 – Hadamard 126
 – modified Allan 126
 – of position solution 8
 – of unit weight 709
 variational equation 939
 vector
 – network analyzer (VNA) 527, 1200
 – tracking 438, 1300
 – tracking architecture 500
 vegetation cover 1179, 1181
 vehicle ad hoc network (VANET) 856
 vehicle-to-infrastructure (V2I) 853
 vehicle-to-vehicle (V2V) 853
 velocity
 – error differential equation 829
 – random walk (VRW) 817
 – relative to the atmosphere 941
 vernal equinox 27, 28, 44, 1300
 vertical
 – alert limit (VAL) 888, 889
 – atmospheric 1121
 – dilution of precision (VDOP) 9, 835, 928
 – ionospheric delay 755
 – protection level (VPL) 353, 920
 – total electron content (VTEC) 178, 567, 728, 756, 953, 974, 1142, 1216, 1222
 very high frequency (VHF) 15, 341, 762, 869, 878, 906, 1021
 – data broadcast (VDB) 883, 906, 1300
 – omnidirectional range (VOR) 341, 878, 1300
 very long baseline interferometry (VLBI) 35, 46, 661, 967, 1040, 1064, 1110, 1164, 1222, 1300
 vibrating beam accelerometer (VBA) 816
 vibrating structure gyroscope (VSG) 815
 vibration rectification error (VRE) 818
 Vienna mapping function (VMF) 176, 729, 986
 virtual
 – clock 300, 1301
 – coherent image 1164
 – reference station (VRS) 775, 776, 778, 1021, 1301
 viscoelastic relaxation 1086
 Viterbi decoder 427, 1301
 volcano deformation 1088

voltage controlled oscillator (VCO) 1189
 voltage standing wave ratio (VSWR) 508
 vulnerability 878, 899

W

WADS 17
 Walker constellation 16, 220, 281, 1301
 warm start 1301
 water vapor profile 1123
 wave propagation 165
 waveform 1165
 wavelength 1301
 wavelength, widelane 671
 waypoint 879
 W-bit 435
 weather
 – forecasting 1115
 – prediction 1128
 weight matrix 8
 weighted least-squares (WLS) 640, 835
 weighting
 – code/phase 999
 – elevation-dependent 995
 Welch's bound 101
 wet
 – delay 1301
 – delay model 173
 – refractivity 170
 – troposphere delay 757
 Whaba's problem 788
 where-in-lane level accuracy 843
 white phase (noise) (WHPH) 124
 Wide Area Augmentation System (WAAS) 14, 185, 216, 340, 343, 762, 846, 883, 922, 1141, 1301
 – master station (WMS) 353
 – reference station (WRS) 353
 wide area GPS enhancement (WAGE) 215
 wide-area DGNSS (WADGNSS) 14, 634
 wide-area DGPS service (WADGPS) 14, 759
 wideband
 – bow-tie turnstile antenna 513
 – interference 476
 WideBandMODEL (WBMOD) 184

- wide-lane 1301
 - combination 586
 - extra 669
 - wavelength 671
 - wifi positioning 851
 - window of acceptance 488
 - wireless local area network (WLAN) 851
 - wireless sensor network (WSN) 856
 - wobble 52
 - World Geodetic System (WGS) 6, 214, 326, 398
 - World Geodetic System 1984 (WGS84) 6, 39, 608, 821, 881, 1301
 - WSCS 98
 - w-test statistic 1301
-
- Y**
-
- yaw angle 86, 571, 785
 - observability 834
 - yaw rate 996
 - yaw steering 85, 570, 1301
 - Y-code 1301
 - year
 - Julian 27
 - tropical 26
-
- Z**
-
- Z-count 424
 - zenith 1301
 - hydrostatic delay (ZHD) 171, 1111
 - total delay 1301
 - troposphere delay (ZTD) 568, 620, 724, 763, 975
 - wet delay (ZWD) 171, 1111
 - zero mean constraint 996
 - zero-baseline 1301
 - single difference 599
 - zero-delay attack 488
 - zero-tide system 1301
 - ZHD
 - uncertainty 1115
 - Z-invariance 668
 - Z-tracking 435, 1301
 - Z-transformation 669, 1301

Recently Published Springer Handbooks

Springer Handbook of Global Navigation Satellite Systems (2017)

ed. by Teunissen, O. Montenbruck, 1328 p., 978-3-319-42926-7

Springer Handbook of Model-Based Science (2017)

ed. by Magnani, Bertolotti, 1157 p., 978-3-319-30525-7

Springer Handbook of Odor (2017)

ed. by Buettner, 1151 p., 978-3-319-26930-6

Springer Handbook of Electrochemical Energy (2017)

ed. by Breitung, Swider-Lyons, 1016 p., 978-3-662-46656-8

Springer Handbook of Robotics (2nd) (2016)

ed. by Siciliano, Khatib, 2227p., 978-3-319-32550-7

Springer Handbook of Ocean Engineering (2016)

ed. by Dhanak, Xiros, 1345 p., 978-3-319-32550-7

Springer Handbook of Computational Intelligence (2015)

ed. by Kacprzyk, Pedrycz, 1633 p., 978-3-662-43505-2

Springer Handbook of Marine Biotechnology (2015)

ed. by Kim, 1512 p., 978-3-642-53970-1

Springer Handbook of Acoustics (2nd) (2015)

ed. by Rossing, 1286 p., 978-1-4939-0754-0

Springer Handbook of Spacetime (2014)

ed. by Ashtekar, Petkov, 887 p., 978-3-642-41991-1

Springer Handbook of Bio-/Neuro-Informatics (2014)

ed. by Kasabov, 1230 p., 978-3-642-30573-3

Springer Handbook of Nanomaterials (2013)

ed. by Vajtai, 1222 p., 978-3-642-20594-1

Springer Handbook of Lasers and Optics (2nd) (2012)

ed. by Träger, 1694 p., 978-3-642-19408-5

Springer Handbook of Geographic Information (2012)

ed. by Kresse, Danko, 1120 p., 978-3-540-72678-4

Springer Handbook of Medical Technology (2011)

ed. by Kramme, Hoffmann, Pozos, 1500 p., 978-3-540-74657-7

Springer Handbook of Metrology and Testing (2nd) (2011)

ed. by Czichos, Saito, Smith, 1229 p., 978-3-642-16640-2

Springer Handbook of Crystal Growth (2010)

ed. by Dhanaraj, Byrappa, Prasad, Dudley, 1816 p., 978-3-540-74182-4

Springer Handbook of Nanotechnology (3rd) (2010)

ed. by Bhushan, 1961 p., 978-3-642-02524-2

Springer Handbook of Automation (2009)

ed. by Nof, 1812 p., 978-3-540-78830-0

Springer Handbook of Mechanical Engineering (2009)

ed. by Grote, Antonsson, 1576 p., 978-3-540-49131-6

Springer Handbook of Experimental Solid Mechanics (2008)

ed. by Sharpe, 1096 p., 978-0-387-26883-5

Springer Handbook of Speech Processing (2007)

ed. by Benesty, Sondhi, Huang, 1176 p., 978-3-540-49125-5

Springer Handbook of Experimental Fluid Mechanics (2007)

ed. by Tropea, Yarin, Foss, 1557 p., 978-3-540-25141-5

Springer Handbook of Electronic and Photonic Materials (2006)

ed. by Kasap, Capper, 1406 p., 978-0-387-26059-4

Springer Handbook of Engineering Statistics (2006)

ed. by Pham, 1120 p., 978-1-85233-806-0

Springer Handbook of Atomic, Molecular, and Optical Physics (2nd) (2005)

ed. by Drake, 1506 p., 978-0-387-20802-2